

montanha viva

Sistema Previsional Inteligente de Suporte à Decisão em Sustentabilidade



T6.6. Disseminação e publicação de artigos em revistas e conferências

Agosto 2025



Conteúdos

Conteúdos	2
Sumário executivo	3
1. Introdução	4
2. Publicações em revistas internacionais	4
3. Publicações em Conferências Internacionais	5
4. Anexos	5

Sumário executivo

O projeto Montanha Viva visa desenvolver um sistema de apoio à decisão, à operacionalidade inteligente e em tempo real na exploração económica das plantas de montanha, especialmente em localizações remotas (sem ligação à internet), com vista a estimular o aproveitamento económico de plantas existentes, o aumento da produção, a redução de consumo de recursos naturais, contribuindo para a promoção da biodiversidade e preservação da sustentabilidade ambiental, em particular, das plantas silvestres de montanha. Partir-se-á da identificação e caracterização de plantas de montanha com características potenciadoras de mitigação natural de pragas e doenças em culturas agrícolas e com propriedades de aplicação em saúde e bem-estar, para a criação de um sistema de sensorização local e remota para análise do vigor das plantas aliado a algoritmos de inteligência artificial para suporte à decisão na realização de atividades culturais em plantas existentes ou em novas explorações agroflorestais. Tem como objetivos:

- Recolher informação de base e produzir conhecimento na identificação e caracterização de plantas de montanha com propriedades de aplicação em saúde e bem-estar e com características potenciadoras de mitigação natural de pragas e doenças em culturas agrícolas na região de montanha da Serra da Gardunha, promovendo a sustentabilidade das explorações agroflorestais existentes e o desenvolvimento de novos produtos e novos negócios a partir do aproveitamento económico da flora silvestre.
- Avaliar e caracterizar as propriedades biológicas de espécies selecionadas com base na recolha de informação a partir de inquéritos etnobotânicos.
- Adaptar soluções tecnológicas existentes e/ou desenvolvimento de soluções específicas para a monitorização local em zonas remotas (sem acesso a fontes de energia elétrica nem a comunicações) e inóspitas (com gradientes termo-higrométricos muito elevados).
- Analisar a potencialidade da deteção remota de alta resolução para determinação em tempo quase-real do vigor das plantas assim como da sua taxa de crescimento.
- Desenvolver um sistema previsional inteligente do vigor de plantas de montanha e de informação e suporte à decisão em sustentabilidade ambiental com vista a otimizar a cultura/exploração das plantas silvestres na região de montanha.
- Promover um conhecimento sustentável, através da instalação de mesas interpretativas e de informação digitais com identificação e divulgação da valia ambiental, paisagística e patrimonial da flora que visam a sensibilização e ordenamento da visita das zonas de montanha.
- Dinamizar trilhos turísticos para a promoção da sustentabilidade da montanha por consciencialização da biodiversidade local.
- Comunicar, divulgar, transferir conhecimento e tecnologia e disseminar os resultados do projeto.

Este documento apresenta os detalhes das publicações científicas em revistas internacionais e conferências internacionais que foram realizadas ao longo da execução física do projeto.

Keywords: Turismo de Montanha, sustentabilidade, publicações, revistas internacionais, conferências internacionais

1. Introdução

Decorrente das atividades de I&D do projeto Montanha Viva, serão publicados na totalidade 11 artigos científicos. Até ao momento encontram-se já publicado 5 artigos em actas de conferências/congressos internacionais e 2 em revistas internacionais com revisão por pares, indexadas às bases científicas SCOPUS e ISI. Para além destes, foram submetidos 4 artigos científicos a revistas internacionais. Abaixo encontra-se a listagem destes artigos, enquanto no anexo se encontram os artigos. As atividades de disseminação e publicação de artigos em revistas e conferências foram muito significativas, denotando o carácter inovador da investigação e desenvolvimento desenvolvida ao longo do projeto.

2. Publicações em revistas internacionais

1. Videira, J., Gaspar, P.D., Soares, V.N.G.J., Caldeira, J.M.L.P. (2023). Detecting and monitoring the development stages of wild flowers and plants using computer vision: Approaches, challenges and opportunities. *International Journal of Advances in Intelligent Informatics*, 9(3), 347-362. (DOI: 10.26555/ijain.v9i3.1012) (Q3; CiteScore=2.600; SJR=0.273; h=14)
2. Videira, J., Gaspar, P.D., Soares, V.N.G.J., Caldeira, J.M.L.P. (2024). A mobile application for detecting and monitoring the development stages of wild flowers and plants. *Informatica – An International Journal of Computing and Informatics*, 48(6), 43-58. (DOI: 10.31449/inf.v48i6.5645) (Q4; CiteScore=6.200; SJR=0.242; h=38)
3. Coimbra, A., Gallardo, E., Luís, A., Gaspar, P.D., Ferreira, S., Duarte, A.P. Phytochemical profiling and bioactivity evaluation of wild plants. *Molecules* (Q1; IF=4.600; CiteScore=8.600; SJR=0.865; h=261) (submetido à revista. Em processo de revisão por pares)
4. Galvão, M., Pereira, N., Gaspar, P.D. Modern Image Segmentation: An Extensive Review. *Image and Vision Computing Image and Vision Computing* (Q1; IF=4.200; CiteScore=7.100; SJR=0.791; h=150) (submetido à revista. Em processo de revisão por pares)
5. Coimbra, A., Luís, Â., Gaspar, P.D., Ferreira, S., Duarte, A.P. Assessment of Antimicrobial Activity of Plant Extracts and Their Synergistic Potential with Conventional Antibiotics Against *Staphylococcus aureus*. *Antibiotics* (Q1; IF=4.600; CiteScore=8.700; SJR=1.114; h=99) (submetido à revista. Em processo de revisão por pares)
6. Sousa, M., Alves, A., Antunes, R., Aguiar, M.A., Gaspar, P.D., Pereira, N. Advancing smart farming and ecological monitoring: Gathering sensing, computational vision, communications technologies and artificial intelligence. *Journal of Sensor and Actuator Network* (Q1, IF=4.200, CiteScore=9.400, SJR=0.875, h=47) (submetido à revista. Em processo de revisão por pares)

3. Publicações em Conferências Internacionais

1. Gaspar, P.D., Lima, T.M., Pombo, J., Duarte, A.P., Monteiro, J., Ferreira, S., Luís, A., Gonçalves, J.C., Neto, P., O'Hara, K., Brás, R., Santos, S. (2023). Uma abordagem integrada à cultura de plantas silvestre e ao turismo em regiões de montanha por via de um sistema previsional inteligente de suporte à decisão em sustentabilidade: Montanha Viva. S4agro International Congress 2023, Cine-Teatro Avenida, Castelo Branco, Portugal, March 2-3, 2023.
2. Coimbra, A., Luís, Â., Gaspar, P.D., Ferreira, S., Duarte, A.P. (2024). Phytochemical characterization and evaluation of antimicrobial properties of wild plants collected in the mountain region of Serra da Gardunha, Portugal. 18th Congress of the International Union of Microbiological Societies (IUMS), Florence, Italy, October 23-25, 2024. (<https://iums2024.com/>)
3. Ferreira, D., Sousa, M., Corceiro, A., Veloso, M., Alves, D., Alves, A.C., Aguiar, M.L., Antunes, R., Pereira, N., Gaspar, P.D. (2024). Technological innovations for remote monitoring and AI-based decision support in mountain ecosystems: The MontanhaViva Project. ICEUBI 2024, Covilhã, Portugal, 27-29, November, 2024.
4. Ferreira, D., Sousa, M., Corceiro, A., Veloso, M., Alves, D., Alves, A.C., Aguiar, M.L., Antunes, R., Pereira, N., Gaspar, P.D. (2024). Implementation of interactive digital panels to promote environmental sustainability and tourism in Serra da Gardunha: The MontanhaViva approach. ICEUBI 2024, Covilhã, Portugal, 27-29, November, 2024.
5. Coimbra A., Luís, Â., Gaspar, P.D., Ferreira, S., Duarte, A.P. (2025). Evaluation of the bioactive activities of wild plants from Serra da Gardunha Mountain region in Portugal. V International Congress in Health Sciences Research 2025: From molecule to community (HSRCongress2025), Covilhã, Portugal, March 20-22, 2025. (<https://www.hsrcongress.pt/>).

4. Anexos

Em seguida são incluídos os artigos referenciados anteriormente.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/374730479>

Detecting and monitoring the development stages of wild flowers and plants using computer vision: approaches, challenges and opportunities

Article in *International Journal of Advances in Intelligent Informatics* · November 2023

DOI: 10.26555/ijain.v9i3.1012

CITATIONS

0

4 authors, including:



Pedro Dinis Gaspar

Universidade da Beira Interior

379 PUBLICATIONS 1,759 CITATIONS

[SEE PROFILE](#)



Vasco N. G. J. Soares

Polytechnic Institute of Castelo Branco

93 PUBLICATIONS 1,332 CITATIONS

[SEE PROFILE](#)



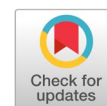
João M. L. P. Caldeira

Polytechnic Institute of Castelo Branco

42 PUBLICATIONS 313 CITATIONS

[SEE PROFILE](#)

Detecting and monitoring the development stages of wild flowers and plants using computer vision: approaches, challenges and opportunities



João Videira ^{a,1}, Pedro Dinis Gaspar ^{b,c,2}, Vasco Nuno da Gama de Jesus Soares ^{a,d,3},
João Manuel Leitão Pires Caldeira ^{a,d,4,*}

^a Polytechnic Institute of Castelo Branco, Av. Pedro Álvares Cabral nº 12, 6000-084 Castelo Branco, Portugal

^b Department of Electromechanical Engineering, University of Beira Interior, Rua Marquês d'Ávila e Bolama, 6201-001 Covilhã, Portugal

^c C-MAST Center for Mechanical and Aerospace Science and Technologies, University of Beira Interior, 6201-001 Covilhã, Portugal

^d Instituto de Telecomunicações, Rua Marquês d'Ávila e Bolama, 6201-001 Covilhã, Portugal

¹ jvideira@ipcbcampus.pt; ² dinis@ubi.pt; ³ vasco.g.soares@ipcb.pt; ⁴ jcaldeira@ipcb.pt

* corresponding author

ARTICLE INFO

Article history

Received February 1, 2023

Revised April 16, 2023

Accepted April 21, 2023

Available online October 3, 2023

Keywords

Wild flowers

Development stages

Computer vision

Machine learning

Deep learning

ABSTRACT

Wild flowers and plants play an important role in protecting biodiversity and providing various ecosystem services. However, some of them are endangered or threatened and are entitled to preservation and protection. This study represents a first step to develop a computer vision system and a supporting mobile app for detecting and monitoring the development stages of wild flowers and plants, aiming to contribute to their preservation. It first introduces the related concepts. Then, surveys related work and categorizes existing solutions presenting their key features, strengths, and limitations. The most promising solutions and techniques are identified. Insights on open issues and research directions in the topic are also provided. This paper paves the way to a wider adoption of recent results in computer vision techniques in this field and for the proposal of a mobile application that uses YOLO convolutional neural networks to detect the stages of development of wild flowers and plants.



This is an open access article under the [CC-BY-SA](#) license.



1. Introduction

Over the last years, farmers, horticulturalists, gardeners, and curious onlookers have been increasingly using new technologies for a more sustainable and efficient agriculture, and to gain experience caring for plants by identifying plant species and diseases. Detecting the stages of the lifecycle of a plant has been used to better understand the stages of growth of the crops to improve efficiency and productivity [1]. But it can be very interesting to apply this concept to wild flowers and plants, as it can help to protect endangered species in protected areas, such as genisteae, nettles or juniper. In some counties, one species of wildflower becomes extinct every two years [2]. Some wild flowers and plants may play a big part on the diet of certain animal species or play an essential role in ecosystems, which may collapse causing the extinction of other species [3]. Furthermore, some of these plants play a key role in the development of modern medications or beauty products [4].

The development stages of wild flowers and plants can be classified as follows: 1) sprout: this stage typically happens underground where the plant starts to grow out of its seed; 2) seedling: this stage is characterized by the spread of roots and the appearance of the first leaves; 3) vegetative: this stage is identified by the development of stems and foliage; 4) budding: this stage can be identified by the

appearance buds on the plant; 5) flowering & pollination: this stage is recognized by the appearance of flowers, which in consequence causes pollination and can be accompanied by the appearance of fruits in early stages; 6) ripening: this stage is identified by the appearance of fruits already matured.

To the best of our knowledge, at the time of this research, no work has been undertaken to specifically detect and monitor the growing stages of wild flowers and plants using computer vision techniques. Computer vision is an application of machine learning and artificial intelligence that takes information from digital images and videos and makes meaningful decisions based on that information. Over the years convolutional neural networks have been widely used for object detection and classification, and various techniques have proven superior results in terms of detection accuracy, speed, objectiveness, reliability.

The work presented in this paper represents a first step in an ongoing effort to develop a system and a support mobile phone app, which on one hand can help park visitors in their enjoyment and awareness of the wild flowers and plants they find along the roadways and trails (flowers they observe, appreciate, and probably photograph), and on the other hand at the same time allows monitoring their development stages. Thus, contributing for wild flowers and plants understanding and preservation.

This paper first introduces the related concepts. Then, presents a survey that focuses on recent peer-reviewed studies (mainly from 2018 to 2022) searched in electronic databases, and existing mobile apps from AppStore and Play Store, guided by three research questions: (a) What types of modern computer vision techniques are commonly used in this area?; (b) What studies and mobile apps have been focused on plant identification?; (c) What studies and mobile apps have been focused on plant disease detection? It aims to identify the most promising approaches to apply to this specific scenario of detecting and monitoring the development stages of wild flowers and plants.

The rest of the paper is organized as follows. Section 2 presents computer vision techniques that have been used in the literature for plant identification and plant disease detection. Section 3 reviews related studies and applications. Section 4 discusses the challenges and provides directions for future developments and research. Finally, Section 5 presents the conclusions and draws some lines of future work.

2. Method

Computer vision makes it possible for systems and computers to take actions or make recommendations based on relevant information gleaned from digital images, videos, and other visual inputs. The field of computer vision has been fundamentally altered by deep learning, which is frequently used to teach computers to “see” and analyse the environment in the same way that people do.

Deep learning involves using a neural network to teach an algorithm through training. With the help of the neural networks, an algorithm can be trained with a large quantity of data and learn from it. Then, it will be able to receive an input such as an image and predict a value for the input based on the data it learned from [5]. Object detection and classification through deep learning can be divided into two tasks. The task of object detection deals with the identification of an object in an image. Thus, in the context of this work, it detects wild flowers and plants in a photo. The task of object classification deals with the categorization of the object based on previously defined classes or types [6]. In the context of this work, it allows determining the wild flowers and plants species and classifying their development stage.

In the last years object detection has made significant progress by using Convolutional Neural Networks (CNN). Due to the capacity of imitating neurons on the brain, CNNs have the characteristic to learn through a large quantity of data [5].

The object detection models that make use of CNNs can be classified in two categories: single-stage and two-stage. The single-stage object detection models produce bounding boxes around the detected object. These bounding boxes are the result of the process of assigning predictions to various regions of the image with the use of anchor boxes. These boxes are used to capture the object and contain a

prediction value [7]. Then, the network will evaluate these predictions and detect the object creating a bounding box around it [8]. Examples of single-stage object detection models are YOLO, KNN, MobileNet and SVM. The two-stage object detection models add a classification stage to the process, which classifies the objects within the subsets of images or regional proposals. This additional stage increases accuracy, although it is slower than single-stage object detection models [8]. Examples of two-stage object detection models are Mask R-CNN and AlexNet.

The main principles of object detection models commonly used for plant identification and plant disease detection are described below.

2.1. Single-stage Object Detection

2.1.1. YOLO

YOLO (You Only Look Once) [9] is a real-time object detection algorithm. YOLO divides the input image into a grid of cells. Each of these cells is responsible for predicting a set of bounding boxes and class probabilities. This allows YOLO to process an entire image in a single pass, hence the name "You Only Look Once" [9].

YOLOv3, the third version of YOLO, uses the CNN Darknet-53 which has 53 convolutional layers for feature extraction. Furthermore, it is the first version of YOLO that can detect objects of different size. This is due to the anchor boxes being capable to be scalable. Consequently, it performs better at the task of object detection. In addition, this algorithm also uses feature pyramid networks (FPN) that allows the algorithm to be able to detect objects at different sizes [10].

YOLOv4 works by receiving an input, which is an image that is passed through three components: the backbone, neck, and head. The backbone is used to extract features by using the CNN CSPDarkNet53 for the feature extraction, which is the process of transforming data into numerical features [11], [12]. The neck is used to extract different features maps from the many stages of the backbone by using Path Aggregation Network (PAN) [11]. Finally, the head, which is composed by dense prediction and sparse prediction [12] is responsible to detect the objects drawing bounding boxes around them [11]. All these components work to produce as result a object surrounded by a bounding box [13]. The Fig. 1 illustrates how this algorithm works.

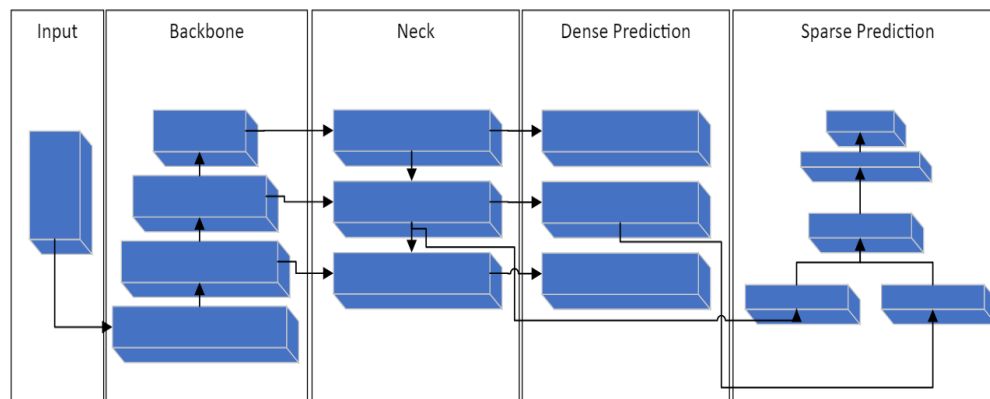


Fig. 1. Architecture of YOLOv4.

2.1.2. KNN algorithm

K-Nearest Neighbours (KNN) [14] algorithm is a classification and regression supervised machine learning technique [15]. This algorithm creates 'K' values by the extraction of features on a training dataset. These 'K' values are then positioned based on their values on a space and are aggregated into classes. After this process, when the algorithm receives an image to classify, it will calculate the 'K' value of the image. Then the 'K' value of the image is used to position the received data in the space. Then, to classify the image, it will compare the distances to the classes of the training dataset by using a formula to calculate the nearest 'K' value [16].

There are other classifier types based on KNN that use different distance metrics, such as: distance weight for Weighted KNN [17] that uses the distance weight; Cubic KNN [18] that uses the cubic distance; the cosine KNN [19] that uses the cosine distance; the Fine KNN [20] that uses the euclidean distance. Fig. 2 illustrates how this algorithm classifies new data.

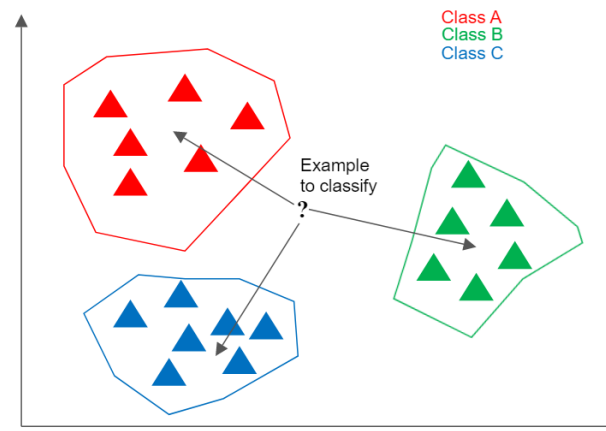


Fig. 2. KNN classifying new data based on the distances to the existing classes.

2.1.3. MobileNet

MobileNet [21] is an accurate and efficient lightweight convolutional neural network (CNN) designed for mobile and embedded devices with limited computational resources. MobileNet uses depthwise separable convolution to reduce the computation required by a traditional CNN. In depthwise separable convolution, a single filter is applied to each channel of the input image. Then, a pointwise convolution is applied to combine the results from each channel. This approach reduces the number of filters applied, and thus reduces the computation required. In addition, this CNN also introduces a technique called width multiplier and resolution multiplier. This technique allows to control the number of channels in the layers and the resolution of the input image, which is a trade-off between computational cost and accuracy.

Furthermore, MobileNet makes use of batch normalization. This is a technique that is used after the depthwise convolution and pointwise convolution layers, to normalize the activations before they are passed through the next layer [22]. By normalizing the activations, the batch normalization helps to reduce the internal covariate shift, which can improve the performance and stability of the network during training [22]. Fig. 3 illustrates how this algorithm works.

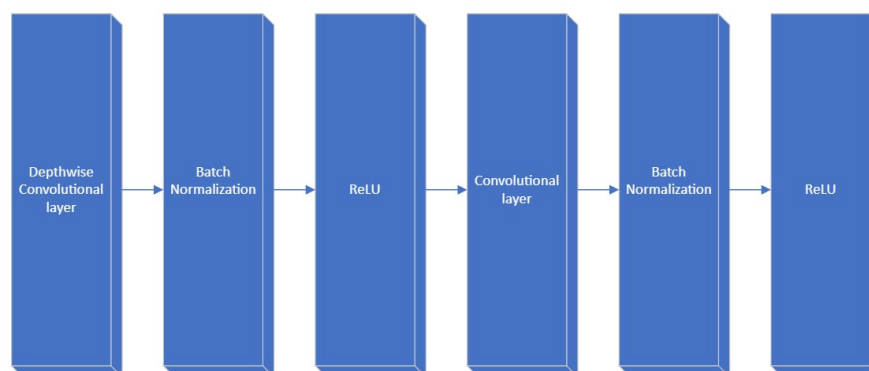


Fig. 3. Architecture of MobileNet.

2.1.4. SVM

Support Vector Machine (SVM) [23] approach consists of mapping data into a high-dimensional feature space so that this data can be categorized. After SVM is trained, it will classify the training data into vectors, which will be placed into a n-dimensional space. Then, a hyperplane will be drawn between

the data categorized with the aid of the selected support vectors, which are the vectors closest to the hyperplane. These support vectors, which are used to represent these data points, form the basis for the SVM method. They are crucial to the algorithm's ability to correctly classify new data.

There can also be multiple hyperplanes to aid the classification of data. Once the hyperplane is drawn, data can be classified by determining on which side of the hyperplane it is placed [24]. Fig. 4 illustrates how this algorithm works.

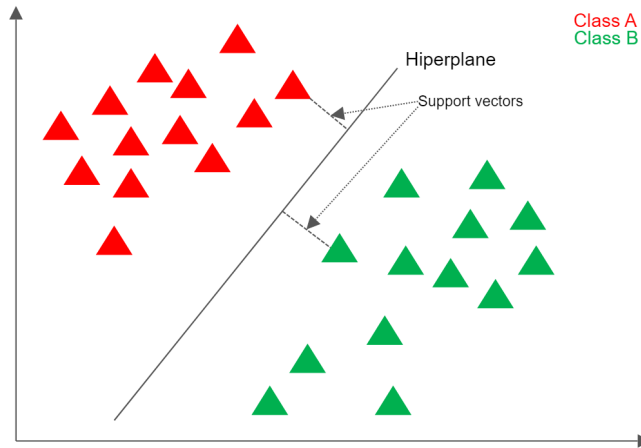


Fig. 4. Vectors divided into two classes by the hyperplane.

2.2. Two-stage Object Detection

2.2.1. Mask R-CNN

Mask R-CNN [25] is based on Faster R-CNN [26]. Faster R-CNN merges Fast R-CNN with a Regional Proposal Network (RPN), making it faster while maintaining its accuracy [27].

Faster R-CNN is a two-stage algorithm [28] that consists of two modules: RPN and Fast R-CNN. RPN is used to generate region proposals. This region proposals are passed to the component Region of Interest (RoI) pooling, which resizes each region proposal to a fixed size before feeding it into the fully connected layers. These layers will then output the class, which is the class label of the object and the bounding box [29].

The algorithm Mask R-CNN differentiates from the Faster R-CNN by replacing the RoI Pooling with RoI align, which is used for extracting a small feature map, and by adding a mask. A new branch was also added to this algorithm that takes regional proposals and inputs them into convolutional layers, which in turn will output a mask [25]. Thus, Mask R-CNN has 3 outputs: the bounding box, the class, and the mask [26]. Fig. 5 illustrates how the Mask R-CNN algorithm works.

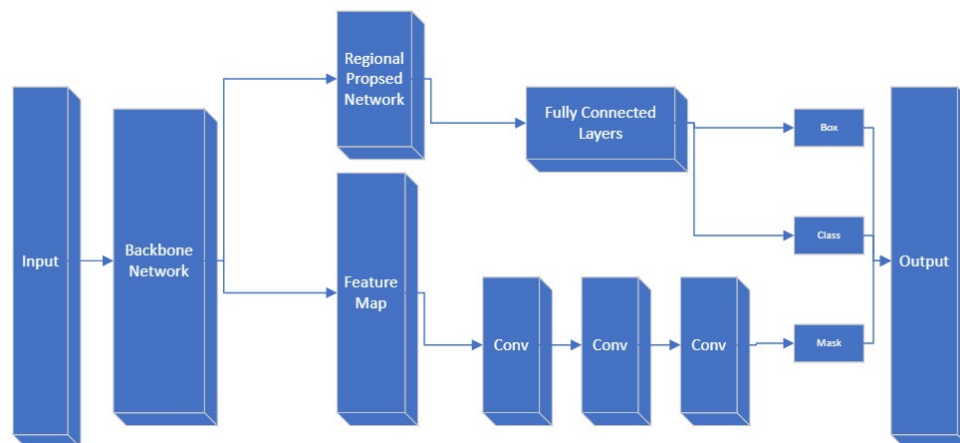


Fig. 5. Architecture of Mask R-CNN.

2.2.2. AlexNet

AlexNet [30] consists of eight layers of which three are fully connected (FC) layers and the other five are convolution layers, with Dropout on the first two layers to reduce overfitting. In addition, the final fully linked layer is followed by a Softmax function, which is used to convert the class scores into a probability distribution over all possible classes. The input picture is filtered by the first convolutional layer, which uses 96 11x11 kernels. The output of the first convolutional layer is filtered by the second convolutional layer, which uses 256 5x5 kernels. The image is then passed to the third, fourth, and finally to the fifth convolutional layers, respectively, each having 384, 384, and 256 3x3 kernels. Then, the output is passed through the first and second fully connected layers with Rectified Linear Unit (ReLU). Finally, the data is passed through the third last fully connected layer or output layer [31]. Fig. 6 illustrates how this algorithm works.

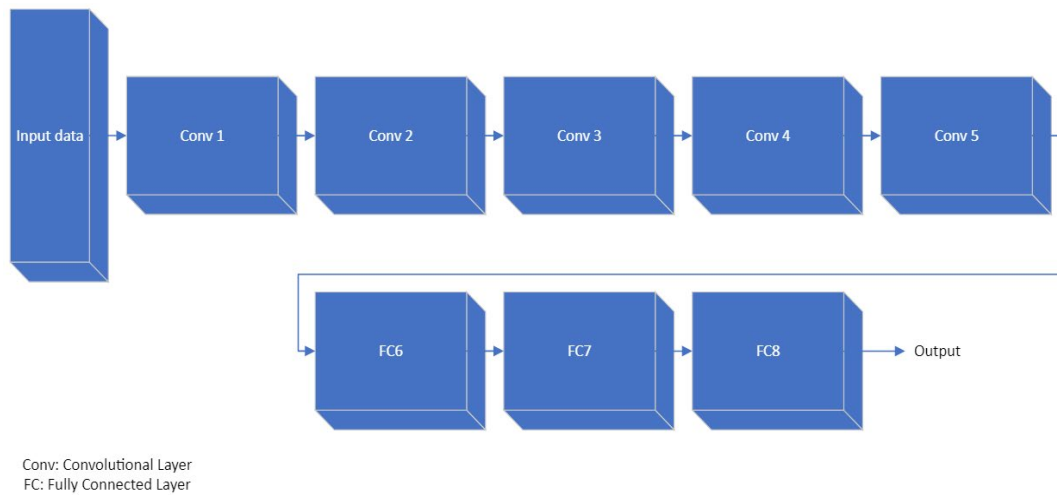


Fig. 6. Architecture of AlexNet.

3. Results and Discussion

To the best of our knowledge, no work to date utilized computer vision techniques to specifically detect and monitor the growing stages of wild flowers and plants. Therefore, the related work presented in this section focuses on recent research works and applications using computer vision techniques for plant identification and plant disease detection.

2.2.3. Plant Identification

Plant detection and identification can be achieved through a photo taken of a plant that is passed through a CNN algorithm, which then gives the output indicating the corresponding plant species. Several mobile applications (apps) available through platforms such as Google Play Store provide this capability. Example of such apps are [32], [33] which additionally present information related to the plant identified. Recent works in the literature have also tackled the same issue. The work in [34] aimed to improve the accuracy and the performance of real-time lemons detection in a natural environment. This improvement was made by switching the backbone of the algorithm YOLOv3 with SE_ResGNet3. The performance of the proposed algorithm was tested and compared with the standard YOLOv3. It was shown that it performs best with an accuracy of 96.28% and 90 frames per second (FPS), while standard YOLOv3 achieved an accuracy of 90.6% and 62 FPS.

Another work [35] had the objective of identifying 15 plant species through their leaves. To achieve this goal, AlexNet was used. It was observed that AlexNet achieved an accuracy of 72%. The work [36] proposed a new algorithm for the identification of poisonous and harmful plants, called Weight Bat-inspired Algorithm (WBA) with Deep Neural Network (DNN). The performance of this algorithm was tested and compared with other algorithms such as SVM, KNN, Naive Bayes (NB), C4.5, Random Forest (RF) and AdaBoost. The proposed algorithm achieved an accuracy of 98% after data

augmentation. Whereas the algorithms SVM, KNN, NB, C4.5, RF and AdaBoost achieved accuracies of 92.7%, 92.7%, 90%, 94.7%, 94% and 94% respectively, after data augmentation.

Other work [37] tried to identify wild plants through their leaves, fruits, or both. Three algorithms were tested, AlexNet, RF, and SVM. It was concluded that AlexNet was the most suitable for the task, achieving the highest accuracy of 98%. The tests also showed that SVM and RF achieved an accuracy of 96.7% and 96% respectively. The work presented in [38] focused in identifying apple flowers in a natural environment. A combination of the algorithms YOLOv4 and channel pruning was considered. The performance of this combined algorithm was compared with Faster R-CNN, Tiny-YOLO v2, YOLO v3, SSD 300 and EfficientDet-D0. It was observed that the combined algorithm performed better than the other algorithms. It achieved a higher accuracy of 97.31%. While the other algorithms Faster R-CNN, Tiny-YOLO v2, YOLO v3, SSD 300 and EfficientDet-D0 obtained the accuracies of 85.10%, 81.75%, 83.12%, 91.64% and 89.52%, respectively. The work [39] had the objective of identifying medicinal plants. It used a variation of KNN called Weighted KNN, which alters the procedures to assign the weights to the 'K' points. This algorithm achieved an accuracy of 98.62%. Another work [40] combined two algorithms, Principal Component Analysis (PCA) and KNN, to identify medicinal plants. PCA was used for feature extraction, and KNN was used for image classification. The results of the study showed that this combination of PCA and KNN achieved an accuracy of 88.67%. In [41], the authors proposed a system to detect peaches using deep learning, based on the algorithm Faster R-CNN. This algorithm was tested, and it achieved an accuracy of 90%. A summary of the above-described applications and related works is reported in Table 1, highlighting their contributions.

Table 1. Approaches/techniques for plant identification.

Name	Type	Year of publication	Types of plants classified	Goal	Technique used	Accuracy
PictureThis - Plant Identifier [32]	App	2017	All types	Identify plants through a picture and deliver relevant information to the user	Unknown	Unknown
NatureID [33]	App	2020	All types	Identify plants or diagnose them through a picture and deliver relevant information to the user	Unknown	Unknown
Lemon-YOLO [34]	Paper	2021	Lemons	Improve the accuracy and the performance of real-time lemons detection in a natural environment	YOLOv3 with SE_ResGNet34 YOLOv3	96,28% 90.60%
Plant Species Image Recognition using Artificial Intelligence on Jetson Nano Computational Platform [35]	Paper	2021	Limited to 15 species	Identify the species of the plant through their leaves	AlexNet	72%
WBA-DNN: A hybrid weight bat algorithm with deep neural network for classification of poisonous and harmful wild plants [36]	Paper	2021	Harmful wild plants	Classify poisonous and harmful plants	WBA-DNN	98%
					SVM	92.7%
					KNN	92.7%
					NB	90%
					C4.5	94.7%
					RF	94%
					AdaBoost	94%

Table 1. (Continued)

Name	Type	Year of publication	Types of plants classified	Goal	Technique used	Accuracy
A New Deep Learning System for Wild Plants Classification and Species Identification: Using Leaves and Fruits [37]	Paper	2022	Wild plants	Identify wild plants through their leaves, fruits, or both	AlexNet	98%
					SVM	96.7%
					RF	96%
Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments [38]	Paper	2020	Apple flowers	Identify apple flowers in a natural environment	YOLOv4 with channel pruning	97.31%
					Faster R-CNN	85.10%
					Tiny-YOLO v2	81.75%
					YOLO v3	83.12%
					SSD 300	91.64%
					EfficientDet-D0	89.52%
Segmentation and identification of medicinal plant through weighted KNN [39]	Paper	2022	Medicinal plants	Identify medicinal plants	Weighted KNN	98.62%
Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants [40]	Paper	2021	10 species of medicinal plants	Identify medicinal plants	Combination of PCA and KNN	88.67%
Peaches Detection Using a Deep Learning Technique A Contribution to Yield Estimation, Resources Management, and Circular Economy [41]	Paper	2022	Peaches	Detection of peaches	Faster R-CNN	90%

2.2.4. Plant Disease Detection

Plant disease detection can also be performed using a photo of a plant that is passed through a CNN algorithm [42]. Some mobile applications available through platforms such as Google Play Store provide this capability [43], [44]. Recent works in the literature show promising results and are described next.

The work presented in [45] aimed to classify rice diseases by using only colour features. It tested different classifiers such as SVM, DC, KNN, NB, DT, RF and Logistic Regression (LR). It was concluded that the SVM classifier presents the best results, achieving an accuracy of 94.65% while DC, KNN, NB, DT, RF and LR obtained the accuracies 92.34%, 91.39%, 75.72%, 83.18%, 92.52%, 75.85% respectively. Another work [46] proposed an algorithm to identify crop diseases. This algorithm makes use of a CNN to extract the features, which are then classified using an error-correcting output codes (ECOC) based on SVM classifier. This algorithm was tested using two CNN's, AlexNet and VGG19. It was concluded that the VGG19 presented the best accuracy of 98.9%. The algorithm with AlexNet achieved an accuracy of 98.8%.

The work in [47] tried to improve the standard AlexNet with Inception-V4 to increase the accuracy of plant disease diagnosis. The proposed algorithm was tested along with AlexNet, VGG11, ZFNet and VGG16. It was concluded that the improved algorithm achieved an accuracy of 96.5%. This accuracy was higher than the algorithms AlexNet, VGG11, ZFNet and VGG16 that obtained an accuracy approximately of 80%.

The work [48] also tested different algorithms to classify plant diseases to determine which has the highest accuracy. These algorithms were AlexNet, SVM Linear and SVM Radial Basis Function (RBF). It was concluded that the most suitable algorithm would be AlexNet. It achieved the highest accuracy of 91.15%, while SVM Linear and RBF achieved accuracies of 88.96% and 89.69% respectively. In [49], the authors aimed to create a technology to detect diseases on tomatoes through their leaves. To achieve this goal the following algorithms were compared - InceptionV3, ResNet50, AlexNet, MobileNetV1, MobileNetV2, MobileNetV3 Large and MobileNetV3 Small. These algorithms were trained and tested with a large range of optimizers, like Adam, Adagrad, SGD and RMSProp. It was concluded that the combination that presented the best accuracy was MobileNetV3 Large with Adagrad, obtaining an accuracy of 99.81%. Additionally, the study also concluded that the best optimizers for each algorithm were MobileNetV3-L with Adagrad, Inception V3 with SGD ResNet50 with SGD, MobileNetV1 with SGD, MobileNetV2 with SGD, MobileNetV3-S with Adagrad and AlexNet with SGD, achieving the accuracies of 99.81%, 99.62%, 99.62%, 99.49%, 98.93%, 98.99% and 96.68% respectively.

The work presented in [50] described a new algorithm to identify plant diseases in crops such as wheat, cotton, grape, corn, and cucumbers. This algorithm combines AlexNet with Particle Swarm Optimizer (PSO). To assess the accuracy of this proposal, the algorithm was tested against AlexNet and the results were compared. It was observed that AlexNet and AlexNet + PSO scored 95.6% and 98.83% respectively.

The work [51] focused on identifying diseases on crops through images of the plants leaves. It used a more compact version of the algorithm YOLOv4, called YOLOv4-tiny. This algorithm was tested and it achieved an accuracy of 63.31%. In [11], two YOLO algorithms, YOLOv3 and YOLOv4, were tested with the goal of identifying seventeen diseases on thirteen plant species. This study concluded that YOLOv4 was the most suitable for the task of detecting diseases on crops. YOLOv3 and YOLOv4 achieved accuracies of 53.08% and 55.45% respectively. The work [52] aimed to create a system capable to identify early blight and late blight on potatoes through their leaves. The paper tested two algorithms to assess which was the most suitable, GoogleNet and AlexNet. Both algorithms achieved an accuracy of 98.51%. Nevertheless, AlexNet outperformed GoogleLeNet in terms of precision 99.44%, sensitivity 98.35%, specificity 99.72%, and F1 score 98.89%. GoogleLeNet achieved a precision of 98.88%, sensitivity 98.34%, specificity 99.44%, and F1 score 98.61%.

Other work [53] tried to identify 7 kinds of diseases that affect strawberries using Mask-R-CNN. It was observed that the algorithm Mask-R-CNN had an accuracy of 82.43%. In [54] a new algorithm was proposed to identify three major leaf diseases on tea plants. This new algorithm is based on Mask R-CNN and wavelet transform, which are then inputted into a four-channeled residual network (F-RNet). This new algorithm was compared with ResNet18, VGG16, AlexNet and SVM to assess its accuracy. The new algorithm achieved an accuracy of 88%. In contrast the algorithms ResNet18, VGG16, AlexNet and SVM registered accuracies of 82%, 80%, 73% and 65% respectively. Another work in [55] proposed the creation of a decision-making support system by classifying the diseases affecting peaches. MobileNetV2 algorithm was trained and tested for the task. It was observed that this algorithm achieved an accuracy of 96%.

A summary of the above-described applications and related works is reported in Table 2, highlighting their contributions.

Table 2. Approaches/techniques for plant disease detection.

Name	Type	Year of publication	Types of plants classified	Purpose of study/app	Technique used	Accuracy
Plant Disease Identification a [43]	App	2020	Fruits, vegetables, and crops	Identify the disease of a specific plant	Unknown	Unknown
Plantix [44]	App	2015	Plants, vegetables, and crops	Identify the disease of a specific plant	Unknown	Unknown
Rice plant disease classification using colour features: a machine learning paradigm [45]	Paper	2020	Rice Plants	Identify rice plant diseases trough machine learning	SVM	94.65%
					DC	92.34%
					KNN	91.39%
					NB	75.72%
					DT	83.18%
					RF	92.52%
Plant Disease Identification and Classification Using Convolutional Neural Network and SVM [46]	Paper	2021	Crops	Identify crop diseases with a new algorithm	LR	75.85%
					VGG19	98.9%
Improved AlexNet with Inception-V4 for Plant Disease Diagnosis [47]	Paper	2022	Crops	Improve AlexNet with Inception-V4 to increase its accuracy for plant disease diagnosis	AlexNet	96.5%
					Inception-V4	80%
					AlexNet	80%
					VGG11	80%
					ZFNet	80%
Comparison of Plant Leaf Classification Using Modified AlexNet and Support Vector Machine [48]	Paper	2021	Nine different plants	Propose an algorithm to classify plant diseases	VGG16	80%
					AlexNet	91.15%
					SVM Linear	88.96%
Optimized Deep Learning Algorithms for Tomato Leaf Disease Detection with Hardware Deployment [49]	Paper	2022	Tomatoes	Determine which algorithm and optimizer is most suitable to detect diseases on tomatoes through their leaves	SVM RBF	89.69%
					MobileNetV3-L with Adagrad	99.81%
					Inception V3 with SGD	99.62%
					ResNet50 with SGD	99.62%
					MobileNetV1 with SGD	99.49%
					MobileNetV2 with SGD	98.93%
					MobileNetV3-S with Adagrad	98.99%
Optimization of Deep Learning Model for Plant Disease Detection Using Particle Swarm Optimizer [50]	Paper	2022	Wheat, cotton, grape, corn, and cucumbers	Detect plant diseases in five types of crops	AlexNet with SGD	96.68%
					AlexNet + PSO	98.83%
Real-Time Detection and Identification of Plant Leaf Diseases using YOLOv4-tiny [51]	Paper	2021	Tomato, mango, strawberry, beans and potato	Identify diseases on crops trough images of the plant leaves	AlexNet	95.6%
					YOLOv4-tiny	63.31%

Table 2. (Continued)

Name	Type	Year of publication	Types of plants classified	Purpose of study/app	Technique used	Accuracy
Plant Disease Detection Based on YOLOv3 and YOLOv4 [11]	Paper	2021	Thirteen plant species	Train two algorithms with YOLOv3 and YOLOv4 to identify seventeen diseases on thirteen plant species	YOLOv4 YOLOv3	55.45% 53.08%
Disease Detection In Plant Leaves Using Deep Learning Models: AlexNet And GoogLeNet [52]	Paper	2021	Potatoes	Create a system to identify early blight and late blight on potatoes through their leaves	AlexNet GoogLeNet	98.51% 98.51%
An Instance Segmentation Model for Strawberry Diseases Based on Mask R-CNN [53]	Paper	2021	Strawberries	Identify 7 kinds of strawberry diseases	Mask R-CNN	82.43%
Symptom recognition of disease and insect damage based on Mask R-CNN, wavelet transform, and F-Rnet [54]	Paper	2022	Tea plants	Propose a new algorithm to identify 3 major leaf diseases of tea plants	Algorithm based on Mask R-CNN, wavelet transform and F-RNet	88% 82% 80% 73% 65%
Decision-making support system for fruit diseases classification using Deep Learning [55]	Paper	2020	Peach	Propose a support system for the classification of peach diseases	ResNet18 VGG16 AlexNet SVM MobileNetV2	96%

4. Challenges and Opportunities

As stated in the introduction the main objective of this ongoing project, is to develop a system, that can help park visitors in their enjoyment and awareness of the wild plants and flowers. The system will be able to identify the plant and flowers species and their growth stages. To achieve this objective the system will be composed by two different components: a mobile application, which can be installed on the mobile phone of each visitor, and fixed devices to be mounted on the ground near the plants and flowers. The mobile application can be used by park visitors themselves. They can direct the camera to a plant or a flower and the application will detect and identify the plant species and growth stage, and provide other relevant information. The fixed devices must be mounted on the ground and equipped with a camera pointing towards the area of wild plants or flowers to be monitored. Each fixed device performs a daily photo capture and sends it to a central computer. Each capture is then classified and kept in a database in the cloud, where researchers, naturalists and enthusiasts can access it. This system will allow constant monitoring of wild flowers and plants which can help in their preservation.

To develop the proposed system four main challenges were identified that need to be overcome. The first challenge is that to the best of our knowledge, at the time of this research, a dataset of images with the growth stages of wild flowers and plants is not available. This represents one of the biggest obstacles, since a dataset is needed to train the algorithm that will be used to classify the flowers and plants development stages.

The second challenge to be considered is the network coverage. This problem occurs when the network signal is weak or even not available in certain areas, which can lead to poor network performance, slow data speeds, and dropped packets. For the mobile application, to avoid large amounts of data being transferred over the network, the classification can be performed directly in the users' mobile phones. Then, the mobile application would only retrieve from the central computer additional information related to the plant and its growth stage. In contrast, the fixed device would transfer the photos to the central computer for classification and storage. Since fixed devices will be mounted in remote areas with high forest density, network coverage may not be the most suitable for transferring large amounts of data. Thus, to carry out these transfers, it would be possible to take advantage of the multiple devices spread across the terrain. A wireless mesh network could be used to bring network coverage to all these devices. Mesh networks create a network of interconnected devices. These devices can relay data between one another, allowing for communication even in areas where there is no direct connection to the destination or to the Internet. This solution would allow photos to be transferred opportunistically via other devices to be delivered to the central computer.

The third challenge identified is related to the power supply for fixed devices. Since these devices would be applied in remote areas, it would be impractical to provide a constant energy supply. Therefore, the operation of these devices would need to rely on a battery. One possible solution would be to use solar panels to recharge devices batteries. However, in locations with high forest density this solution may not be suitable due to lack of sunlight. To mitigate this problem, some mechanisms could be used to minimize the operation of the devices (e.g., use duty cycles).

Finally, the fourth challenge would be the possible destruction of the fixed devices by wild animals. Wildlife may disturb the fixed device by tumbling it to the ground or changing the camera angle. As a possible solution, it would be advisable to install the system at some distance from the ground, out of reach of wild animals. Alternatively, camouflage could be used over the devices to avoid detection by wild animals.

The development of this system will also create some opportunities. Three identified opportunities are described below. One of the opportunities identified is that, to the best of our knowledge at the time of the development of this work, a technology has not yet been developed that allows the classification of the growth stage of a particular wild plant through a photo. Therefore, the creation of this system would contribute greatly for the preservation of endangered species of wild flowers and plants.

The second opportunity would be the improvement of pre-existing technologies, which are used for the identification of plants or plant diseases. An indication of this opportunity was shown in the related work section. According to a study [50] using the App Plantix it was proven that only 10% of the infected plants had the right disease diagnosed. Despite this bad result, it shouldn't be assumed that all Apps have a bad accuracy, but it is a clue that some results may not be as accurate as expected.

The third opportunity is that the proposed system would provide information to researchers and park workers to help in the protection of specific wild flowers and plants species. In addition, it would also provide information and awareness to the public (i.e., park visitors) through the mobile application.

The proposed system will classify and identify the growth stages of wild flowers and plants on a central computer by using computer vision techniques. The analysis of the survey presented in this paper allows to conclude that for the fixed device the most suitable algorithms would be MobileNetV3-Large with the optimizer Adagrad or AlexNet. As presented in [49], using MobileNetV3-Large with the Adagrad optimising algorithm, an accuracy of 99.81% was achieved. In [37], [48], [52] and [35] it was

proven that the AlexNet algorithm is also very reliable and accurate in many situations. Therefore, AlexNet can also be considered.

Based on the results obtained in [12], [34] and [11], YOLOv4 or YOLOv3 proved to be the best option to be used for the mobile application. These algorithms are fast, use relatively low processing power, and can perform results in real-time. These features fit perfectly with the requirements of the proposed mobile application.

5. Conclusion

Not only are wild plants and flowers beautiful, but they are also an important part of our lives. They feed both people and animals or can be used for aromatic or medical purposes. Wildlife like bees, birds, butterflies, and others would not exist without them. However, this irreplaceable natural heritage is in danger of being lost due to human activity and climate change. The work presented in this paper contributes to the conservation effort. It identified computer vision as a suitable technological platform for detecting and monitoring the development stages of wild flowers and plants. An overview of the most used computer vision techniques used in this area was provided. Then, a survey of plant identification and plant disease detection related research and applications was presented. It aimed to identify the most promising computer vision techniques. From the literature review, several open issues and challenges in the area were presented. The research that is presented in this paper is a first step in an ongoing effort to create a system and a mobile App that will support park visitors in their enjoyment and awareness of the wild flowers and plants they find along the roadways and trails. It will also allow the remote monitoring and collecting information regarding wild flowers and plants development stages. We are currently evaluating and comparing the precision of the above-identified computer vision techniques on different datasets of wild flowers and plants.

Acknowledgment

J.M.L.P.C. and V.N.G.J.S. acknowledge that this work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020. P.D.G. thanks the support provided by the Center for Mechanical and Aerospace Science and Technologies (C-MAST) under project UIDB/00151/2020.

This is within the activities of project Montanha Viva – An intelligent prediction system for decision support in sustainability, project PD21-00009, promoted by PROMOVE program funded by Fundação La Caixa and supported by Fundação para a Ciência e a Tecnologia and BPI.

Declarations

Author contribution. Conceptualization, P.D.G., J.V; methodology, J.V; validation, P.D.G., J.M.L.P.C. and V.N.G.J.S.; formal analysis, P.D.G., J.M.L.P.C. and V.N.G.J.S.; investigation, J.V; writing—original draft preparation, J.V; writing—review and editing, P.D.G., J.M.L.P.C. and V.N.G.J.S.; supervision, J.M.L.P.C. and V.N.G.J.S.; funding acquisition, P.D.G., J.M.L.P.C. and V.N.G.J.S. All authors have read and agreed to the published version of the manuscript.

Conflict of interest. The authors declare no conflict of interest.

References

- [1] A. Wang, W. Zhang, and X. Wei, “A review on weed detection using ground-based machine vision and image processing techniques,” *Comput. Electron. Agric.*, vol. 158, pp. 226–240, Mar. 2019, doi: [10.1016/j.compag.2019.02.005](https://doi.org/10.1016/j.compag.2019.02.005).
- [2] T. Dines, “Plantlife – A Voice for Wildflowers,” *ArkWildlife*. Accessed Jan. 04, 2023. [Online]. Available at: <https://www.arkwildlife.co.uk/blog/plantlife-a-voice-for-wildflowers/>.
- [3] “Chain-reaction extinctions will cascade through nature: Study,” *Daily Sabah*, 2022. Accessed Feb. 04, 2023. [Online]. Available at: <https://www.dailysabah.com/life/environment/chain-reaction-extinctions-will-cascade-through-nature-study>.

- [4] E. V. Christaki and P. C. Florou-Paneri, "Aloe vera: A plant for many uses," *J. Food, Agric. Environ.*, vol. 8, no. 2, pp. 245–249, 2010, [Online]. Available at: https://www.researchgate.net/publication/265268175_Aloe_vera_A_plant_for_many_uses.
- [5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: [10.1038/nature14539](https://doi.org/10.1038/nature14539).
- [6] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, "Contextualizing object detection and classification," in *CVPR 2011*, Jun. 2011, pp. 1585–1592, doi: [10.1109/CVPR.2011.5995330](https://doi.org/10.1109/CVPR.2011.5995330).
- [7] "Anchor Boxes for Object Detection," *MATLAB & Simulink*. Accessed Jan. 06, 2022. [Online]. Available at: <https://www.mathworks.com/help/vision/ug/anchor-boxes-for-object-detection.html>.
- [8] "What Is Object Detection?," *MATLAB & Simulink*. Accessed Jan. 06, 2022. [Online]. Available at: https://www.mathworks.com/discovery/object-detection.html?s_tid=srchtitle_object_detection_1.
- [9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 779–788, doi: [10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91).
- [10] J. Redmon and A. Farhadi, "YOLO: Real-Time Object Detection." Accessed Jan. 06, 2023. [Online]. Available at: <https://pjreddie.com/darknet/yolo/>.
- [11] A. Shill and M. A. Rahman, "Plant Disease Detection Based on YOLOv3 and YOLOv4," in *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)*, Jul. 2021, pp. 1–6, doi: [10.1109/ACMI53878.2021.9528179](https://doi.org/10.1109/ACMI53878.2021.9528179).
- [12] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *Comput. Vis. Pattern Recognit.*, pp. 1–17, Apr. 2020, Accessed: Jun. 06, 2023. [Online]. Available: <https://arxiv.org/abs/2004.10934v1>.
- [13] M. Valente, H. Silva, J. Caldeira, V. Soares, and P. Gaspar, "Detection of Waste Containers Using Computer Vision," *Appl. Syst. Innov.*, vol. 2, no. 1, p. 11, Mar. 2019, doi: [10.3390/asi2010011](https://doi.org/10.3390/asi2010011).
- [14] O. Kramer, "K-Nearest Neighbors," in *Dimensionality Reduction with Unsupervised Nearest Neighbors*, Springer, Berlin, Heidelberg, 2013, pp. 13–23, doi: [10.1007/978-3-642-38652-7_2](https://doi.org/10.1007/978-3-642-38652-7_2).
- [15] K. Pavani and P. Sriramya, "Comparison of KNN, ANN, CNN and YOLO algorithms for detecting the accurate traffic flow and build an Intelligent Transportation System," in *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, Feb. 2022, pp. 628–633, doi: [10.1109/ICIPTM54933.2022.9753900](https://doi.org/10.1109/ICIPTM54933.2022.9753900).
- [16] D. S. Shakya, "Analysis of Artificial Intelligence based Image Classification Techniques," *J. Innov. Image Process.*, vol. 2, no. 1, pp. 44–54, 2020, doi: [10.36548/jiip.2020.1.005](https://doi.org/10.36548/jiip.2020.1.005).
- [17] H. Yigit, "A weighting approach for KNN classifier," in *2013 International Conference on Electronics, Computer and Computation (ICECCO)*, Nov. 2013, pp. 228–231, doi: [10.1109/ICECCO.2013.6718270](https://doi.org/10.1109/ICECCO.2013.6718270).
- [18] A. F. Abate, M. Nappi, S. Barra, and M. De Marsico, "What are you doing while answering your smartphone?," in *2018 24th International Conference on Pattern Recognition (ICPR)*, Aug. 2018, vol. 2018–August, pp. 3120–3125, doi: [10.1109/ICPR.2018.8545797](https://doi.org/10.1109/ICPR.2018.8545797).
- [19] D. C. Anastasiu and G. Karypis, "Fast Parallel Cosine K-Nearest Neighbor Graph Construction," in *2016 6th Workshop on Irregular Applications: Architecture and Algorithms (IA3)*, Nov. 2016, pp. 50–53, doi: [10.1109/IA3.2016.013](https://doi.org/10.1109/IA3.2016.013).
- [20] Y. Xu, Q. Zhu, Z. Fan, M. Qiu, Y. Chen, and H. Liu, "Coarse to fine K nearest neighbor classifier," *Pattern Recognit. Lett.*, vol. 34, no. 9, pp. 980–986, Jul. 2013, doi: [10.1016/j.patrec.2013.01.028](https://doi.org/10.1016/j.patrec.2013.01.028).
- [21] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *Comput. Vis. Pattern Recognit.*, pp. 1–9, Apr. 2017, Accessed: Jan. 06, 2023. [Online]. Available at: <https://arxiv.org/abs/1704.04861v1>.
- [22] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, pp. 448–456, Feb. 2015, Accessed: Jan. 06, 2023. [Online]. Available at: <https://arxiv.org/abs/1502.03167v3>.

- [23] S. Suthaharan, "Support Vector Machine," in *Machine Learning Models and Algorithms for Big Data Classification*, Springer, Boston, MA, 2016, pp. 207–235, doi: [10.1007/978-1-4899-7641-3_9](https://doi.org/10.1007/978-1-4899-7641-3_9).
- [24] "How SVM Works," *IBM Documentation*, 2021. Accessed Jan. 06, 2022. [Online]. Available at: <https://www.ibm.com/docs/en/spss-modeler/saas?topic=models-how-svm-works>.
- [25] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, vol. 2017-Octob, pp. 2980–2988, doi: [10.1109/ICCV.2017.322](https://doi.org/10.1109/ICCV.2017.322).
- [26] H. He *et al.*, "Mask R-CNN based automated identification and extraction of oil well sites," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 112, p. 102875, Aug. 2022, doi: [10.1016/j.jag.2022.102875](https://doi.org/10.1016/j.jag.2022.102875).
- [27] S. T. Cynthia, K. M. Shahrukh Hossain, M. N. Hasan, M. Asaduzzaman, and A. K. Das, "Automated Detection of Plant Diseases Using Image Processing and Faster R-CNN Algorithm," in *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, Dec. 2019, pp. 1–5, doi: [10.1109/STI47673.2019.9068092](https://doi.org/10.1109/STI47673.2019.9068092).
- [28] A. Lohia, K. D. Kadam, R. R. Joshi, and A. M. Bongale, "Bibliometric Analysis of One-stage and Two-stage Object Detection," *Libr. Philos. Pract.*, vol. 2021, no. February, pp. 1–33, 2021, [Online]. Available at: https://www.researchgate.net/profile/Rahul-Joshi-9/publication/349297260_.
- [29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: [10.1109/TPAMI.2016.2577031](https://doi.org/10.1109/TPAMI.2016.2577031).
- [30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: [10.1145/3065386](https://doi.org/10.1145/3065386).
- [31] M. Pak and S. Kim, "A review of deep learning in image recognition," in *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, Aug. 2017, vol. 2018-Janua, pp. 1–3, doi: [10.1109/CAIPT.2017.8320684](https://doi.org/10.1109/CAIPT.2017.8320684).
- [32] "PictureThis Identificar Planta," *Apps no Google Play*. Accessed Jan. 06, 2023. [Online]. Available at: https://play.google.com/store/apps/details?id=cn.danatech.xingseus&hl=pt_PT&gl=US&pli=1.
- [33] "NatureID - Identificar plantas," *Apps no Google Play*. Accessed Jan. 06, 2023. [Online]. Available at: https://play.google.com/store/apps/details?id=plant.identification.flower.tree.leaf.identifier.identify.cat.dog.breed.nature&hl=pt_PT&gl=US.
- [34] G. Li, X. Huang, J. Ai, Z. Yi, and W. Xie, "Lemon-YOLO: An efficient object detection method for lemons in the natural environment," *IET Image Process.*, vol. 15, no. 9, pp. 1998–2009, Jul. 2021, doi: [10.1049/ipr2.12171](https://doi.org/10.1049/ipr2.12171).
- [35] S. Chavan, J. Ford, X. Yu, and J. Saniie, "Plant Species Image Recognition using Artificial Intelligence on Jetson Nano Computational Platform," in *2021 IEEE International Conference on Electro Information Technology (EIT)*, May 2021, vol. 2021-May, pp. 350–354, doi: [10.1109/EIT51626.2021.9491893](https://doi.org/10.1109/EIT51626.2021.9491893).
- [36] M. H. IBRAHIM, "WBA-DNN: A hybrid weight bat algorithm with deep neural network for classification of poisonous and harmful wild plants," *Comput. Electron. Agric.*, vol. 190, p. 106478, Nov. 2021, doi: [10.1016/J.COMPAG.2021.106478](https://doi.org/10.1016/J.COMPAG.2021.106478).
- [37] N. M. A. Ibrahim, D. G. Gabr, and A.-H. M. Emara, "A New Deep Learning System for Wild Plants Classification and Species Identification: Using Leaves and Fruits," in *Lecture Notes on Data Engineering and Communications Technologies*, vol. 127, Springer Science and Business Media Deutschland GmbH, 2022, pp. 26–37, doi: [10.1007/978-3-030-98741-1_3](https://doi.org/10.1007/978-3-030-98741-1_3).
- [38] D. Wu, S. Lv, M. Jiang, and H. Song, "Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments," *Comput. Electron. Agric.*, vol. 178, p. 105742, Nov. 2020, doi: [10.1016/j.compag.2020.105742](https://doi.org/10.1016/j.compag.2020.105742).
- [39] S. Patil and M. Sasikala, "Segmentation and identification of medicinal plant through weighted KNN," *Multimed. Tools Appl.*, vol. 82, no. 2, pp. 2805–2819, Jan. 2023, doi: [10.1007/S11042-022-13201-7/METRICS](https://doi.org/10.1007/S11042-022-13201-7/METRICS).
- [40] R. I. Borman, R. Napianto, N. Nugroho, D. Pasha, Y. Rahmanto, and Y. E. Pratama Yudoutomo, "Implementation of PCA and KNN Algorithms in the Classification of Indonesian Medicinal Plants," in

- 2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), Oct. 2021, pp. 46–50, doi: [10.1109/ICOMITEE53461.2021.9650176](https://doi.org/10.1109/ICOMITEE53461.2021.9650176).
- [41] E. T. Assunção *et al.*, “Peaches Detection Using a Deep Learning Technique—A Contribution to Yield Estimation, Resources Management, and Circular Economy,” *Climate*, vol. 10, no. 2, p. 11, Jan. 2022, doi: [10.3390/cli10020011](https://doi.org/10.3390/cli10020011).
- [42] V. Tiwari, R. C. Joshi, and M. K. Dutta, “Dense convolutional neural networks based multiclass plant disease detection and classification using leaf images,” *Ecol. Inform.*, vol. 63, p. 101289, Jul. 2021, doi: [10.1016/j.ecoinf.2021.101289](https://doi.org/10.1016/j.ecoinf.2021.101289).
- [43] “Plant Disease Identification a,” *Apps no Google Play*. Accessed Jan. 06, 2023. [Online]. Available at: https://play.google.com/store/apps/details?id=com.fouxa.plantdiseasedetection&hl=pt_PT&gl=US.
- [44] “Plantix - seu médico agrícola,” *Apps no Google Play*. Accessed Jun. 06, 2023. [Online]. Available at: https://play.google.com/store/apps/details?id=com.peat.GartenBank&hl=pt_PT&gl=US.
- [45] V. K. Shrivastava and M. K. Pradhan, “Rice plant disease classification using color features: a machine learning paradigm,” *J. Plant Pathol.*, vol. 103, no. 1, pp. 17–26, Feb. 2021, doi: [10.1007/s42161-020-00683-3](https://doi.org/10.1007/s42161-020-00683-3).
- [46] H. Kibriya, I. Abdullah, and A. Nasrullah, “Plant Disease Identification and Classification Using Convolutional Neural Network and SVM,” in *2021 International Conference on Frontiers of Information Technology (FIT)*, Dec. 2021, pp. 264–268, doi: [10.1109/FIT53504.2021.00056](https://doi.org/10.1109/FIT53504.2021.00056).
- [47] Z. Li *et al.*, “Improved AlexNet with Inception-V4 for Plant Disease Diagnosis,” *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–12, Sep. 2022, doi: [10.1155/2022/5862600](https://doi.org/10.1155/2022/5862600).
- [48] S. A. Wagle and H. R., “Comparison of Plant Leaf Classification Using Modified AlexNet and Support Vector Machine,” *Trait. du Signal*, vol. 38, no. 1, pp. 79–87, Feb. 2021, doi: [10.18280/ts.380108](https://doi.org/10.18280/ts.380108).
- [49] H. Tarek, H. Aly, S. Eisa, and M. Abul-Soud, “Optimized Deep Learning Algorithms for Tomato Leaf Disease Detection with Hardware Deployment,” *Electronics*, vol. 11, no. 1, p. 140, Jan. 2022, doi: [10.3390/electronics11010140](https://doi.org/10.3390/electronics11010140).
- [50] A. Elaraby, W. Hamdy, and M. Alruwaili, “Optimization of Deep Learning Model for Plant Disease Detection Using Particle Swarm Optimizer,” *Comput. Mater. Contin.*, vol. 71, no. 2, pp. 4019–4031, Dec. 2022, doi: [10.32604/cmc.2022.022161](https://doi.org/10.32604/cmc.2022.022161).
- [51] A. Mohandas, M. S. Anjali, and U. Rahul Varma, “Real-Time Detection and Identification of Plant Leaf Diseases using YOLOv4-tiny,” in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2021, pp. 1–5, doi: [10.1109/ICCCNT51525.2021.9579783](https://doi.org/10.1109/ICCCNT51525.2021.9579783).
- [52] O. Saxena, S. Agrawal, and S. Silakari, “Disease Detection In Plant Leaves Using Deep Learning Models: AlexNet And GoogLeNet,” in *2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES)*, Dec. 2021, pp. 1–6, doi: [10.1109/TRIBES52498.2021.9751620](https://doi.org/10.1109/TRIBES52498.2021.9751620).
- [53] U. Afzaal, B. Bhattarai, Y. R. Pandeya, and J. Lee, “An Instance Segmentation Model for Strawberry Diseases Based on Mask R-CNN,” *Sensors*, vol. 21, no. 19, p. 6565, Sep. 2021, doi: [10.3390/s21196565](https://doi.org/10.3390/s21196565).
- [54] H. Li *et al.*, “Symptom recognition of disease and insect damage based on Mask R-CNN, wavelet transform, and F-RNet,” *Front. Plant Sci.*, vol. 13, Jul. 2022, pp. 1–14, doi: [10.3389/fpls.2022.922797](https://doi.org/10.3389/fpls.2022.922797).
- [55] E. Assuncao, C. Diniz, P. D. Gaspar, and H. Proenca, “Decision-making support system for fruit diseases classification using Deep Learning,” in *2020 International Conference on Decision Aid Sciences and Application (DASA)*, Nov. 2020, pp. 652–656, doi: [10.1109/DASA51403.2020.9317219](https://doi.org/10.1109/DASA51403.2020.9317219).

A Mobile Application for Detecting and Monitoring the Development Stages of Wild Flowers and Plants

João Videira ¹, Pedro D. Gaspar ^{2,3}, Vasco N. G. J. Soares ^{1,4*} and João M. L. P. Caldeira ^{1,4}

¹Polytechnic Institute of Castelo Branco, Av. Pedro Álvares Cabral n° 12, 6000-084 Castelo Branco, Portugal

²Department of Electromechanical Engineering, University of Beira Interior, Rua Marquês d'Ávila e Bolama, 6201-001 Covilhã, Portugal

³C-MAST Center for Mechanical and Aerospace Science and Technologies, University of Beira Interior, 6201-001 Covilhã, Portugal

⁴Instituto de Telecomunicações, Rua Marquês d'Ávila e Bolama, 6201-001 Covilhã, Portugal

jvideira@ipcbcampus.pt, dinis@ubi.pt, vasco.g.soares@ipcb.pt, jcaldeira@ipcb.pt

*Corresponding author: g.soares@ipcb.pt

Keywords: wild flowers and plants, development stages, computer vision, convolutional neural networks, YOLOv4, YOLOv4-tiny, mobile app

Received: Januar 7, 2024

Wild flowers and plants appear spontaneously. They form the ecological basis on which life depends. They play a fundamental role in the regeneration of natural life and the balance of ecological systems. However, this irreplaceable natural heritage is at risk of being lost due to human activity and climate change. The work presented in this paper contributes to the conservation effort. It is based on a previous study by the same authors, which identified computer vision as a suitable technological platform for detecting and monitoring the development stages of wild flowers and plants. It describes the process of developing a mobile application that uses YOLOv4 and YOLOv4-tiny convolutional neural networks to detect the stages of development of wild flowers and plants. This application could be used by visitors in a nature park to provide information and raise awareness about the wild flowers and plants they find along the roads and trails.

Povzetek: Raziskava uvaja mobilno aplikacijo z uporabo konvolucijskih nevronskih mrež YOLOv4 za prepoznavanje razvojnih stopenj divjih rastlin, prispevajoč k ohranjanju naravne dediščine.

1 Introduction

Plants are recognized as a vital part of the world's biological diversity and an essential resource for the planet. They play a fundamental role in maintaining basic ecosystem functions and are indispensable for the survival of animal life on our planet [1]. While agricultural plants provide food and basic fibers, many wild plants are of great economic and cultural importance and have enormous potential, serving as food, medicine, fuel, clothing and common shelter [2], [3].

Given the growing number of endogenous and/or wild flowers and plants at risk of extinction and in decline due to climate change and the impact of human action [4], [5], there is an urgent need to contribute technological solutions for their conservation and preservation. Detecting, monitoring and following the development stages of endogenous and/or wild flowers and plants can alert biologists and researchers linked to the various areas of biodiversity to possible problems with the surrounding environment, which can help them make more informed decisions about how to manage and protect natural parks or preserved areas. On the other hand, it can support and help inform tourists and the community in general about wild flowers and plants found along roads and trails. This

promotion of awareness about wild flowers and plants is aimed at promoting environmental sustainability, but also the development of economic and cultural activities in regions where these plants grow.

The development stages of wild flowers and plants can be classified as follows [6]: 1) sprout: this stage typically takes place underground, where the plant begins to grow from its seed; 2) seedling: this stage is characterized by the spread of roots and the appearance of the first leaves; 3) vegetative: this stage is identified by the development of stems and foliage; 4) budding: this stage can be identified by the appearance of buds on the plant; 5) flowering: this stage is recognized by the appearance of flowers, which consequently causes pollination and can be accompanied by the appearance of fruit in the early stages; 6) ripening: this stage is identified by the appearance of ripe fruit. The sprout stage cannot be identified by computer vision techniques, as it takes place underground.

This work follows on from the conclusions presented in a previous paper by the same authors [7], which analyzed computer vision as a suitable technological platform for detecting and monitoring the development stages of wild flowers and plants. It presented a survey of the research in this field and applications related to plant identification and plant disease detection. The most

promising computer vision techniques were identified, and open problems and challenges were discussed.

The work presented in this article is one of the stages of an ongoing project which aims to develop a mobile application and system based on computer vision techniques to detect and monitor the stages of development of wild flowers and plants. The application can be used by visitors to a nature park to provide information and raise awareness about the wild flowers and plants they encounter along the roads and trails. The system will allow scientists and biologists, or the merely curious, to remotely monitor and collect information on the stages of development of wild flowers and plants.

Although there are already mobile applications that can identify plant species, such as those available in [8], [9], to the best of the authors' knowledge, at the time of writing this article, there is still no mobile application capable of classifying the developmental stages of wild flowers and plants. Therefore, the main contributions of this article are: 1) the creation of a dataset with the stages of development of a wild plant; 2) a performance evaluation study of the convolutional neural network (CNN) models YOLOv4 and YOLOv4-tiny for detecting the stages of development of this wild plant; 3) the description of the process of developing a mobile application, compatible with the Android platform, which uses these CNNs.

This application is aimed at visitors or workers in nature parks. It works by capturing an image of the plant or flower. Then, using these computer vision techniques, it will be able to identify the species of the wild plant, as well as determine its stage of development, and to provide additional information about them.

The rest of the paper is organized as follows. Section 2 introduces the main concepts of the computer vision techniques used in the context of this work and presents a performance assessment. Section 3 describes the implementation process and the operation of the mobile application developed in the context of this work. Finally, Section 4 concludes the article and presents future work.

1 Computer vision techniques

Computer vision techniques include a variety of algorithms, models and procedures that allow computers to analyze visual data such as photographs and videos [10] and perform tasks such as object detection and classification [11]. Object detection is the task of locating an object in the visual input, while object classification involves assigning a classification to the objects detected in that same input [12]. These concepts are illustrated in Figure 1. Although these tasks differ, deep learning is often used for both.

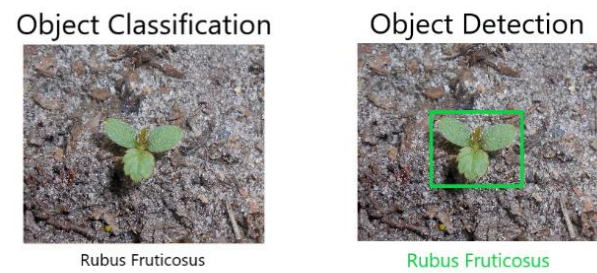


Figure 1: Illustration of object classification and detection concepts.

Deep learning is a subfield of machine learning, which focuses on the creation and training of convolutional neural networks. This training is carried out by learning patterns from a large volume of data [13], such as a dataset of images. This training is possible because CNNs are made up of layers of interconnected nodes, which simulate the behavior of neurons in a human brain [13]. These concepts are illustrated in Figure 2. Training CNNs makes it possible to detect patterns, and consequently to detect and classify objects [14].

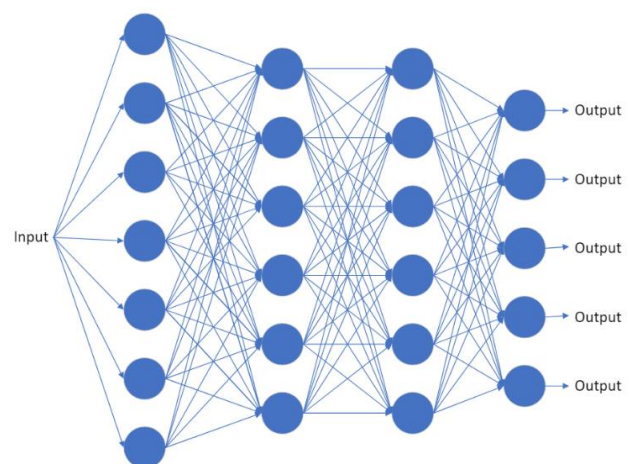


Figure 2: The structure of a convolutional neural network.

In a previous work [7] by the same authors of this article, it was found that over the years, various CNN models have been considered in the literature for the tasks of plant identification and plant disease detection. Given the results reported in [15], [16] and [17], it was concluded that YOLOv3 or YOLOv4 would be the best option to consider when developing a mobile application to detect the development stages of wild flowers and plants. These models are fast, require relatively little processing capacity and allow results to be obtained in real time. These characteristics fit perfectly with the requirements of the mobile application presented in this paper.

1.1 YOLOv4 e YOLOv4-tiny

The YOLO (You Only Look Once) model [18] analyses images quickly by dividing an image into a grid, predicting the bounding boxes, confidence levels and class probabilities of the objects. The result is a set of objects bounding boxes, with class names and confidence levels [18]. These concepts are illustrated in Figure 3.

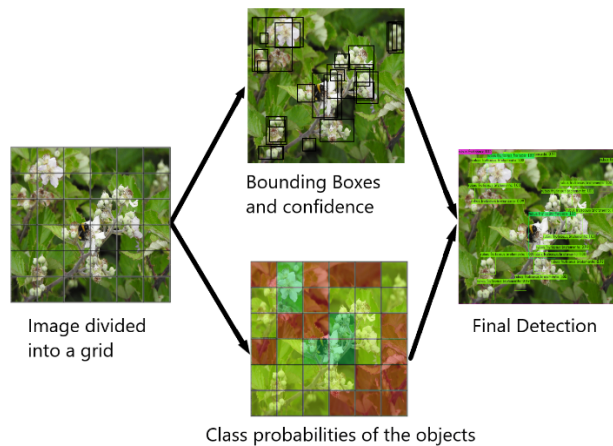


Figure 2: YOLO model detection process.

The YOLOv4 model consists of 3 components: backbone, neck and head. YOLOv4 uses CNN CSPDarkNet53 in the backbone, while YOLOv3 uses DarkNet53 [19]. This component is responsible for feature extraction, which is the process of transforming data into numerical values. In YOLOv4, the neck component uses Path Aggregation Network (PAN) [17] to extract feature maps, while YOLOv3 uses Feature Pyramid Extraction (FPN). Finally, the head component consists of applying anchor boxes to the feature map extracted by PAN. These anchor boxes are used to capture the objects and contain a prediction value [20]. At this stage three heads can be used to identify objects of various sizes, after the feature maps of various scales are joined and subjected to a convolution operation [21].

The YOLOv4-tiny model is a simplified version of YOLOv4 [22]. The first difference in this model is the use of the CSPDarknet53-tiny CNN [23]. The neck component of YOLOv4-tiny uses the Feature Pyramid Network (FPN) structure [23], a design that improves object detection accuracy and increases detection speed [24]. Another difference from YOLOv4 is that YOLOv4-tiny uses only two heads instead of three [23]. This modification could potentially pose challenges when detecting objects at extreme scales, such as very small objects [22]. Despite this limitation, the integration of CSPDarknet53-tiny and FPN into YOLOv4-tiny contributes to its overall performance, allowing it to perform object detection tasks with less computing power and greater speed.

1.2 Performance evaluation

This subsection focuses on evaluating the performance of the YOLOv4 and YOLOv4-tiny CNN models for detecting the developmental stages of wild

flowers and plants. To this end, the dataset created as part of this work, the benchmark scenario, the performance metrics considered, and the experimental results are presented next.

1.2.1 Dataset description

To the best of the authors' knowledge, there are no available datasets with images of the developmental stage's wild flowers and plants. Therefore, it was necessary to create a dataset with the developmental stages of a specific wild plant in order to train and test the models. The plant selected for proof of concept and testing was *Rubus Fruticosus*, also known as "bramble" or "blackberry". This choice was because it is a common wild plant and thus with many images available.

The dataset was categorized into 6 different classes, each representing a stage of development of the wild plant. The initial stage, called seedling, marks the appearance of roots and the initial appearance of leaves. This is followed by the vegetative stage, characterized by the development of stems and foliage. The budding stage shows the appearance of buds on the plant, heralding the next stage of flowering. The flowering stage is identified by the presence of flowers, often accompanied by the appearance of the first fruits. The ripening stage is characterized by the appearance of fruit on the plant. Finally, the class "*Rubus Fruticosus*" is used to identify this plant species. All the images used in this dataset were obtained from the INaturalist [25] and biodiversity4all [26], platforms, taking advantage of the large number of images present on them. This dataset can be found on the Kaggle platform in the following link [27]. Figure 4 shows an example of each of the developmental stages described.

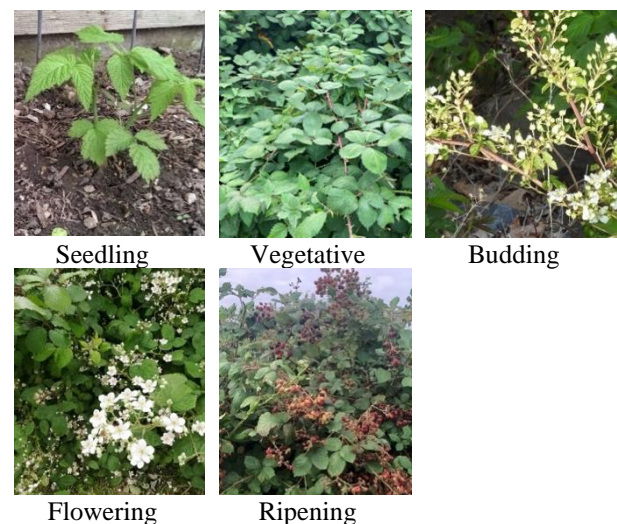


Figure 4: The development stages of the wild plant *Rubus Fruticosus*.

To train and validate the CNN models, the dataset was separated into train and test. In the training dataset, each development stage category was created with approximately 100 images. An exception was the "Budding" category, which, given the limited availability of images, only contains 80 images for training. It is

important to note that the “*Rubus Fruticosus*” class has 490 images because this is the total number of images of all the development stages for this plant species.

To create the test dataset, 20 images were used for each development stage. This distribution of the number of images for training and testing each development stage guarantees sufficient data for training the models and evaluating their performance. Table 1 shows the number of images available for training and testing for each class (i.e., development stage).

Table 1: Number of images for each class (i.e., development stage) in the training and test dataset.

Class	Train	Test
Seedling	100	20
Vegetative	100	20
Budding	80	20
Flowering	110	20
Ripening	100	20
<i>Rubus Fruticosus</i>	490	100

Next, the Yolo_Label tool [28] was used to create the annotations for each image. This task tells the models the location of the objects and their classification. Then, they can then be trained with this data and their performance can be evaluated. Figure 5 illustrates this task.

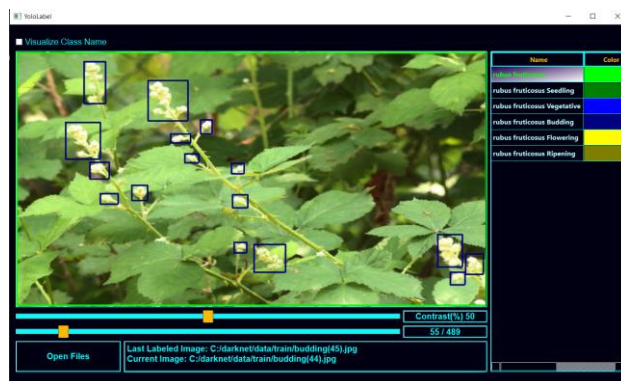


Fig. 3. The process of creating an annotation on a dataset image with the Yolo_Label tool.

1.2.2 Benchmark scenario

Both models were trained and tested on the same device. To do this, the darknet code [29] was downloaded to carry out the necessary training and tests. The test environment was hosted on a device with an AMD Ryzen 5 4600H CPU, 16 GB of RAM, and an NVIDIA GeForce GTX 1650 GPU which increased the computing power required for deep learning operations.

This test environment allowed an unbiased comparison of the accuracy, processing speed and efficiency of the YOLOv4 and YOLOv4-tiny models.

1.2.3 Performance metrics

To assess the performance of YOLOv4 and YOLOv4-tiny in the task of object detection and classification, the

models were trained with 12,000 iterations with batches of 64 images, following the instructions in [29]. This number of iterations is equivalent to approximately 1567 epochs, according to the formula shown in (1).

$$Epochs = \frac{\text{number of iterations}}{\frac{\text{number of images in train}}{\text{batch}}} \quad (1)$$

Once the training was complete, the trained models with the best average precision (mAP) were selected. This metric is calculated according to the formula shown in (2) and considers the accuracy of each class (AP_k) and the number of classes (n). The mAP is the metric commonly used to compare the performance of CNN models such as YOLO.

$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k \quad (2)$$

It is also interesting to analyze in CNN models, the average loss obtained by calculating the average of the total loss of several batches in a dataset, and the test time required to determine the mAP.

Overfitting [30] is a challenge when training CNNs. It occurs when a model becomes excessively specialized in the training data it has been exposed to, to the point of memorizing the details of the data. As a result, the model performs remarkably well on the training data, but fails when confronted with new and unseen data [30]. To try to solve this problem, it is essential to find a balance between accurately capturing significant patterns and avoiding an overly complex model that adapts too much to the training data. Figure 6 illustrates the concept of overfitting.



Figure 4: Illustration of the concept of overfitting.

The solution found to avoid overfitting was the Early Stopping [31]. This solution consists of monitoring the model's performance with a validation dataset while it is being trained. As training is carried out, signs of performance degradation or stagnation are looked for. When it is detected that the model's performance on the validation dataset is no longer improving or is getting worse, which indicates a greater number of errors, the training is stopped. This prevents the model from specializing on the training dataset and helps to ensure that the model maintains good performance against new data

[32]. With this solution, the training of both models was stopped when they showed a degradation or stagnation in performance. Figure 7 illustrates this concept.

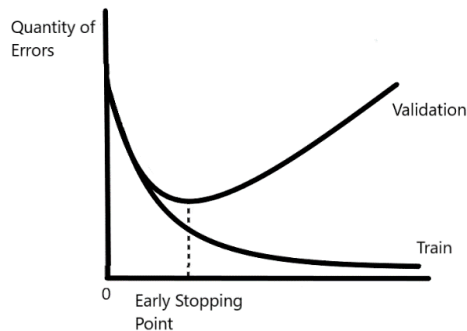


Figure 5: Illustration of the early stopping approach.

1.2.4 Results and discussion

The YOLOv4 and YOLOv4-tiny models went through 12000 training iterations. The model weights were then selected based on the maximum mAP value achieved, for use in the mobile application presented in the next section.

Figure 8 shows the results of the YOLOv4 model training process. The final mAP and average loss values were 74,83% and 1,4794, respectively. It was also concluded that the model achieved its best performance at around 7600 iterations with an accuracy of 77,18%. This performance demonstrates the model's effectiveness in identifying and classifying objects.

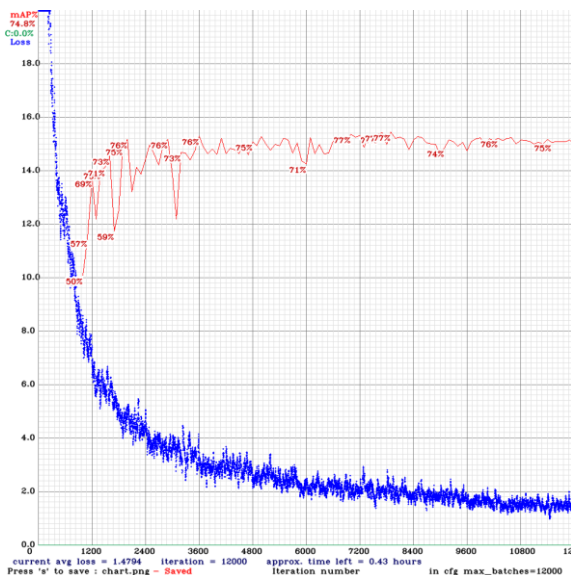


Figure 6: Results of the YOLOv4 training process.

To further analyze the performance of the YOLOv4 model, the accuracy of each class (i.e., development stage) was evaluated. Table 2 shows the accuracy results recorded for each class in the YOLOv4 model. Although each class exceeded the average accuracy, it was observed that the "Ripening" class had a lower average accuracy

with a value of 26,24%. This suggests possible areas of future improvement for object recognition in this specific class.

Table 2: Accuracy results for each class (i.e., development stage) in the YOLOv4 and YOLOv4-tiny models.

Class	mAP	
	YOLOv4	YOLOv4-tiny
<i>Rubus Fruticosus</i>	92,52%	86,03%
Seedling	82,13%	69,86%
Vegetative	100%	72,40%
Budding	82,71%	74,99%
Flowering	79,44%	83,51%
Ripening	26,24%	31,50%

Figure 9 shows the results of the training process for the YOLOv4-tiny model. The model finished its training with a mAP value of 65,43% and an average loss of 0,3566. However, the model achieved its best performance at around 7000 iterations with an accuracy of 69,72%. Therefore, it can be concluded that this model is very effective at classifying and detecting objects, even though it is a simplified version of the YOLOv4 model.

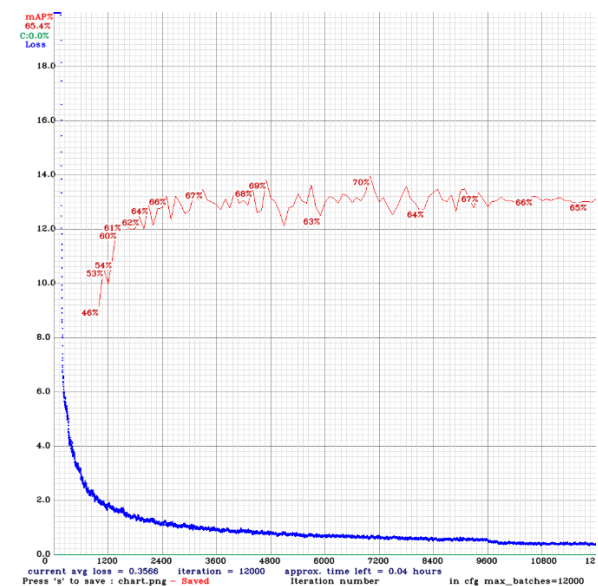


Figure 7: Results of the YOLOv4-tiny training process.

The performance of the YOLOv4-tiny model was also analyzed in depth. Table 2 also shows the accuracy results recorded for each class in this model. It was found that, although most of the classes exceeded the average accuracy, the "Ripening" development stage again showed a considerably lower average accuracy of 31,50%. This result reinforces the conclusion that both models have greater difficulty in recognizing this stage of development.

In view of these results, an explanation was sought as to why the "Ripening" class had a manifestly lower performance when compared to the other classes in both models. The visual appearance of the fruits of this wild plant, which can be red or black depending on the stage

they are at, is how this development stage is detected. This variation in the appearance of the fruit adds complexity and may certainly play an important role in the lower average accuracy value recorded.

Therefore, it is believed that compiling a dataset with a greater number and variety of photographs that capture the various stages of fruit development can help resolve this constraint and contribute to improving model performance. This will improve the ability to recognize and generalize patterns associated with the different appearances of this fruit, resulting in greater accuracy in detecting this stage.

The performance evaluation also showed that the YOLOv4 model completed the test to determine the mAP in 5 seconds, i.e., it detected all the test data in 5 seconds, while the YOLOv4-tiny model completed the same test in 1 second. Figure 10 shows the time and mAP results of the tests carried out.

```
IoU threshold = 50 %, used Area-Under-Curve for each unique Recall
mean average precision (mAP@0.50) = 0.771755, or 77.18 %
Total Detection Time: 5 Seconds

IoU threshold = 50 %, used Area-Under-Curve for each unique Recall
mean average precision (mAP@0.50) = 0.697157, or 69.72 %
Total Detection Time: 1 Seconds
```

Figure 10: Accuracy results and total detection and classification times for YOLOv4 and YOLOv4-tiny respectively.

This difference in test time results suggests that the YOLOv4-tiny model requires less computing power than YOLOv4. YOLOv4-tiny may sacrifice some detection accuracy in favor of processing speed. But the faster processing of the test dataset demonstrates its suitability for scenarios where fast response and real-time performance are key, and/or there are limited computing resources.

2 Mobile application

This section describes the development of the mobile application for Android called "MontanhaVivaApp". It can be used by visitors in a nature park to interactively promote knowledge about the development stages of different wild flowers and plants found along the roads and trails, contributing to their conservation and preservation.

2.1 Methodology

To develop this application, the User-Centred Design and Iterative Development methodologies were adopted. User-Centered Design [33] gives priority to end users throughout the development process, to understand their needs and preferences. This involves techniques such as usability testing to ensure that the resulting product is in line with user expectations. Iterative Development [34] emphasizes continuous improvement, through repeated cycles of design, implementation, and evaluation.

The User-Centered Design methodology was applied to the creation of a prototype in Adobe XD [35]. It enabled usability tests to be carried out on the mobile application's

interface, to find and correct shortcomings, and to determine missing features. The usability testing phase is described in subsection 3.4.

The Iterative Development methodology was applied in the development phase. The project was divided into several reasonably sized tasks, each focusing on a different set of functionalities. The continuous integration of new code and testing of new features ensured that the development went smoothly. This development phase is explained in subsection 3.5.

2.2 Requirement analysis

User requirements refer to the needs and expectations of end users, determining the desired features and interactions [36]. Requirements can be divided into two categories: functional requirements and non-functional requirements.

Functional requirements specify the precise tasks that an application must perform, i.e., the main functionalities and user interactions. To assist in the process of identifying functional requirements, a use case diagram was drawn, shown in Figure 11 in Unified Modeling Language (UML) notation [37]. Use cases represent the interactions between users and the application. They provide a means of capturing the system's requirements and identify the functionalities to which the user has access.

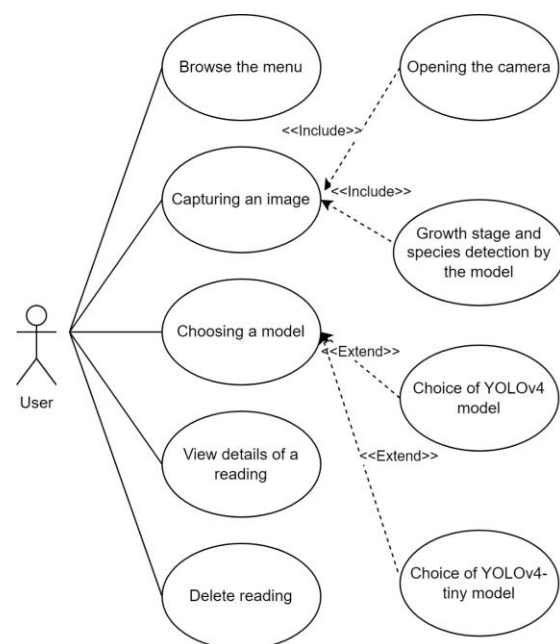


Figure 8: The use case diagram of the application.

Next, the functional requirements are described with the name of the interaction, a description, its preconditions, and outputs. Table 3 summarizes this information.

1) Navigation through the main menu, allowing the user to view previous readings. Its precondition is that the application is initialized. As an output of the requirement, the user can use all the functions found in the main menu,

such as taking a photo to take a reading or checking the details of a previous reading.

2) Capturing a photo to be analyzed by the CNN model. The precondition for this requirement is that the user clicks on the button available in the main menu to take a photo. As an output of the requirement, the image is processed by the chosen model.

3) The choice of CNN template, which can be selected from the top section of the main menu. The precondition for this requirement is that the user is in the main menu. The output of the requirement is that the user can use all the functions found in the main menu.

4) The use of the YOLOv4 model. The precondition for this requirement is that the user has selected this CNN model in the main menu and has taken a photograph. As an output of the requirement, the user returns to the main menu where the result of the new reading is now available.

5) The use of the YOLOv4-tiny template. The precondition for this requirement is that the user has selected this CNN model from the main menu and captured a photo. As an output of the requirement, the user returns to the main menu where the result of the new reading is now available.

6) Seeing all the details of a reading. The precondition for this requirement is that the user has selected one of the readings in the main menu. To exit the requirement, the user can delete the reading or return to the main menu.

7) Delete a specific reading using a button on the details page of a reading. This requirement has the precondition of being on the details page of a reading. As an exit from the requirement, the user returns to the main menu.

The application also has non-functional requirements, which include attributes such as performance, feedback on errors and ease of use.

An intuitive interface refers to the ease of use and interaction of the user interface. It denotes the system's ability to facilitate user involvement and navigation, without the need for extensive training or guidance. It contributes to the application's ease of use by minimizing the learning curve, allowing users to quickly understand the interface's functionalities and access them effortlessly.

Performance is the application's ability to perform tasks effectively and efficiently, even under varying conditions and workloads. It includes factors such as responsiveness, speed and resource utilization. A well-performing application meets user expectations by providing quick responses, fast data processing and smooth functionality across different devices and usage scenarios.

Error feedback refers to the application's ability to give clear and informative answers to users when errors or exceptions occur during its operation. This requirement underlines the importance of maintaining a user-friendly environment, even in the presence of unforeseen problems. Effective error feedback provides users with concise and understandable explanations of the problems encountered, suggests potential solutions and guides them towards troubleshooting or making informed decisions. By ensuring informative error feedback, the application facilitates user understanding, minimizes frustration and promotes a positive user experience.

A scalable database refers to the importance of a structured and adaptable database architecture to which information about other species of wild flowers and plants and their development stages can be added, while maintaining its efficiency and the organization of the data.

These requirements combined form a framework for developing an application that aligns with user expectations, provides the desired functionalities and meets quality expectations.

In addition to the user requirements, the platform requirements must also be addressed. As this application was developed using Android Studio [38], it requires an Android operating system. More specifically, as minSdkVersion 24 was used, the application works on Android versions 7 and later. Installation requires approximately 80 MB of storage, and at least 2 GB of RAM is recommended for optimum performance. A working camera is also essential for capturing photographs. In addition to these requirements, users who choose to use YOLOv4 need an Internet connection to use it.

Table 3: Summary of functional requirements.

Name	Precondition	Exit
Browse the menu	Initialize the application.	Use of any functionality found in the menu.
Capturing an image	Click on the button to capture a photo.	Return to the menu.
Choosing a model	Find yourself in the menu.	Use of any functionality found in the menu.
Choice of YOLOv4 model	Find yourself in the menu.	Use of any functionality found in the menu.
Choice of YOLOv4-tiny model	Find yourself in the menu.	Use of any functionality found in the menu.
View details of a reading	Click on one of the readings.	Delete reading. Return to the menu.
Delete reading	Finding yourself on the details page of a reading.	Return to the menu.

2.3 Technologies and architecture of the mobile app

The "MontanhaVivaApp" mobile application was implemented using the Android Studio integrated development environment (IDE) (electric eel 2022.1.1) [38] and the Java programming language. The application uses an SQLite database [39] and the Structured Query Language (SQL). The database was designed to hold a

large amount of data on the different species of wild flowers and plants and their development stages.

The application uses the YOLOv4 and YOLOv4-tiny trained models to detect and classify images of wild flowers and plants in order to identify the species and their development stage. The YOLOv4-tiny model was implemented locally in the application, as it requires less computing power, allowing it to be used by a mobile device. The YOLOv4 model, which requires more computing power, was implemented remotely and is used via an application programming interface (API). To create this API, the Flask platform in version 2.3.3 [40] was used, together with Python language. Figure 12 shows the diagram of the architecture described.

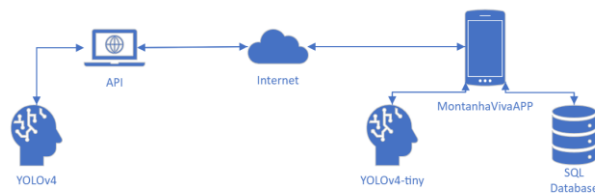


Figure 9: The MontanhaVivaAPP application architecture.

Since the application will be used by visitors in nature parks, they may be in remote locations where Internet access may be scarce or non-existent. Therefore, it was decided to implement the YOLOv4-tiny model and the database locally. This way, even if a user is in an area without network coverage and captures a photograph, the application will be able to identify the wild plant and its development stage, providing a reading with the plant's information. Figure 13 shows a sequence diagram describing the process of creating a new reading locally, using the YOLOv4-tiny model.

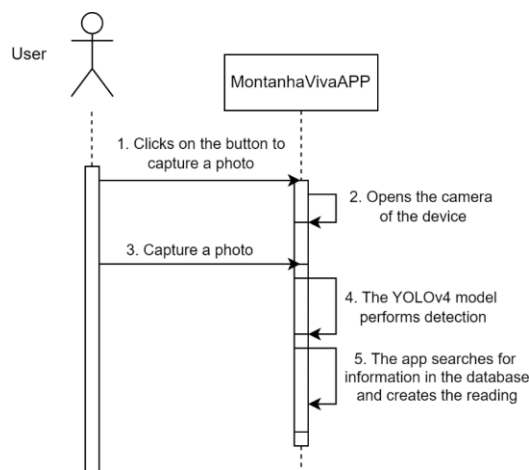


Fig. 10. The sequence diagram illustrating the process of creating a new reading locally using the YOLOv4-tiny model.

In addition, the application offers another option for detection and classification using the YOLOv4 model, for users who have access to the Internet. The user may prefer to use this model because it is more accurate, as discussed in a subsection above. This model is available to the user via an API, which will receive a photograph and use the YOLOv4 model to perform the detection and classification, returning the results to the application (i.e., species and development stage). Then, the application will create a new reading from this data. Figure 14 shows a sequence diagram illustrating the process of creating a new reading locally, using the YOLOv4 model.

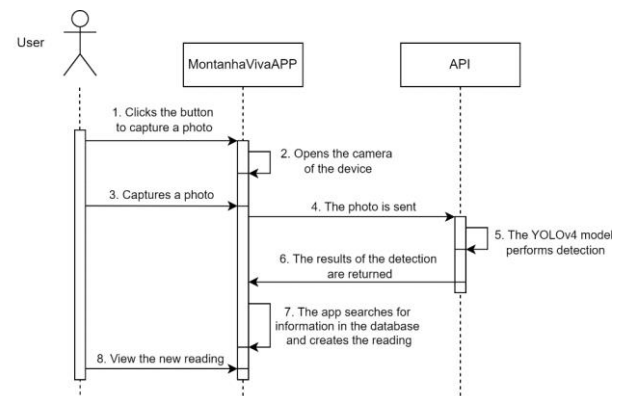


Figure 11: The sequence diagram illustrating the process of creating a new reading using the API with the YOLOv4 model.

2.4 Usability tests

Before starting the application development process, an Adobe XD prototype was developed to simulate the application's functionalities. This prototype is available at [41]. The prototype made it possible to analyze how users would interact with the application, as well as to identify missing features to improve the application.

The initial interface of the prototype is the main menu, which serves as the entry point to the application. This menu contains all the readings of wild plants previously taken by the user. These readings are separated into cells. Each cell displays a summary set of information, including the scientific name, the common name, the stage of development, the period of that stage and the date of the reading. The main menu also has a camera button in the bottom right-hand corner, which simulates the process of capturing a photograph of a wild plant for detection and classification. After successfully capturing the photo, the user returns to the main menu interface, where the newly generated reading allows further exploration and management. Figure 15 shows the process of creating a new reading and shows these described interfaces.

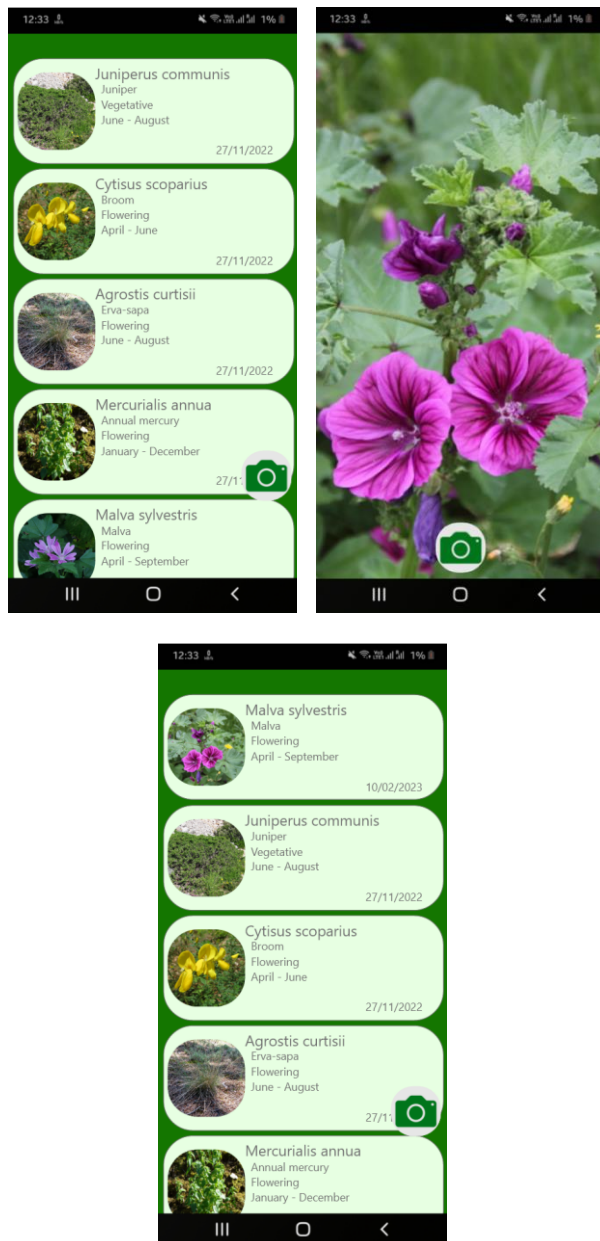


Figure 12: The process of creating a new reading in the prototype.

An interface is also available that allows users to obtain more detailed information about a particular wild plant reading. As it can be seen in Figure 16, it displays information such as the common name, the scientific name, the development stage, the period of the

development stage, the habitat specifications, and a full description of the species.



Figure 13: The interface details of a reading in the prototype.

Usability tests were carried out with real users to obtain information about interaction with the application's interface and the user experience. The concept and purpose of the mobile application were explained, and users were asked to create a new reading and consult its details. The difficulty of performing these tasks was assessed. Users were then asked two questions: Question 1 "Would you use this application?"; Question 2 "Would you add any functionality or information?". Table 4 summarizes the results of the usability tests.

The feedback from the usability tests allowed assessing possible changes to be made to the application. In view of what was reported, a "Delete reading" button was incorporated, giving users control over their readings. The category in which the plant is classified by the International Union for Conservation of Nature (IUCN), was also added to the detailed information, which reflects its degree of risk of extinction.

It was also decided to give autonomy to the user to choose the CNN model to be used for detection and classification, unlike the previous approach which decided automatically based on the existence of network coverage. These adjustments based on user experience were aimed at improving the usability of the application.

Table 4: Usability test results

ID of the user	Problems with the interface	Question 1 Would you use this application?	Question 2 Would you add any functionality or information?
1	None observed.	Yes, I like hiking and it would be good to identify plants and what stage they're at.	I would add a button to choose whether I want to use local detection, as I may not want to waste mobile data.
2	None observed.	No, I'm not in the habit of visiting nature parks.	No.
3	None observed.	Yes, because I have a garden at home.	Add a description of the plant's growth stage.
4	None observed.	Yes, it looks like something I would use on a visit to a nature park.	Add information about the state of danger the plant is in.
5	None observed.	No, I'm not interested in plants.	Add a button to delete a reading.
6	None observed.	Yes, if I went to a nature park.	The option to choose the model so as not to use mobile data.
7	None observed.	Yes, I thought it was an interesting idea.	No.
8	None observed.	Yes, if you visit a nature park.	No.
9	None observed.	No, I'm not in the habit of visiting nature parks.	No.
10	None observed.	Yes, I'm in the habit of hiking and I think it's an interesting thing to use.	Add a description of the growth stage.

2.5 Development

The Iterative Development methodology was used to develop the mobile application. Thus, the process was

divided into several tasks. Initially, all the interfaces and menus were created and the navigation between them was tested, ensuring that the new code was implemented in this iteration without any problems.

Next, the SQLite database implemented in the application was created. At this stage, the application only returns results for the *Rubus Fruticosus* species, due to the difficulties when creating the dataset described in a subsection above. However, both the application and the database have been designed to store and display information on the development stages of various plant species.

The database stores the information that will be shown to the user. It consists of 6 tables, shown in the entity-relationship (ER) model in Figure 17, which store information such as: the common name of the species, the scientific name of the species, the development stage, the time period of the development stage, the IUCN code, the habitats in which it can be found, and descriptions of the species and stage. After planning and implementing the database, its operation was tested.

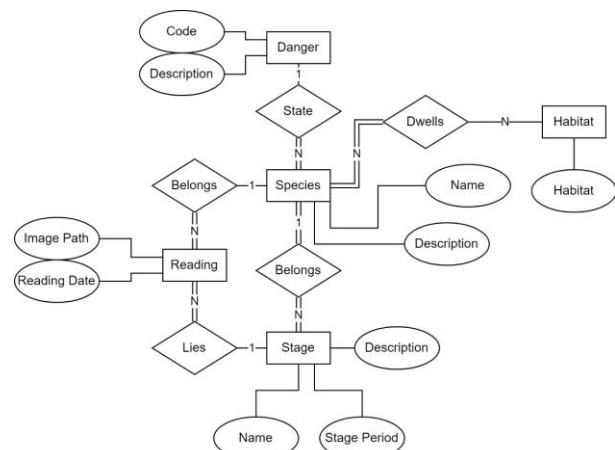


Figure 14: The ER model of the database.

After implementing the database, the YOLOv4-tiny model was implemented. To do this, it was necessary to convert the model trained in darknet to tensorflow lite [42]. This approach allowed the model to be implemented locally in the application. After this implementation, the model was rigorously tested with all the development stages to ensure that any errors were detected and corrected.

Next, an API was implemented with the Flask framework to allow the YOLOv4 model to be executed remotely. To do this, the application checks if there is access to the API using a mobile Internet connection (3G or higher). If successful, it sends the photo for detection and classification. The results are returned in JSON format, and include the species and development stage, to form the reading that is then displayed in the application. The API was subjected to functional tests to detect and correct errors. Figure 18 illustrates how this works.

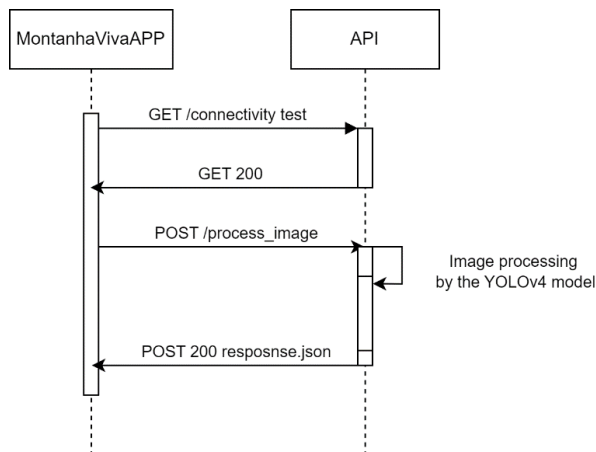


Figure 15: The sequence diagram illustrating the interaction between the application and the API.

After these iterations, the functional prototype of the MontanhaVivaApp application was developed and is available at [43]. Figure 19 shows the modules of the mobile application, including the Java files and menus, respecting the Java language nomenclature [44].

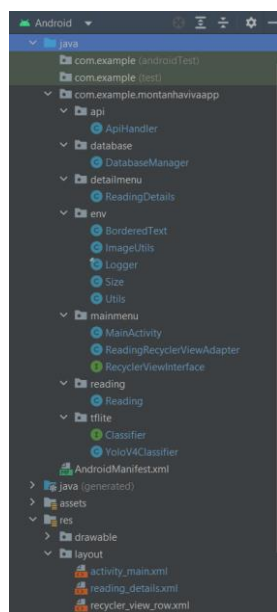


Figure 16: The modules of the MontanhaVivaApp mobile application.

2.6 Structure of the application and evaluation

This subsection describes the features of the MontanhaVivaApp application and the main operations that can be performed on it.

The main menu, shown in Figure 20, serves as the entry point to the application. It provides a view of the previously recorded wild flowers and plants readings. These readings are listed in cells in the main menu, each containing summary information about the scientific name, the common name, the development stage, the period of the stage and the date of the reading.

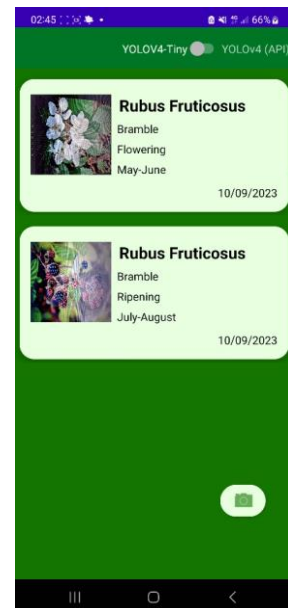


Figure 17: The main menu interface.

Users can consult more detailed information about a particular reading. All they need to do is to select one of the readings listed in the main menu and the interface shown in Figure 21 will appear. This screen provides a range of information including the common name of the plant, the scientific name, the development stage, the period of that stage, the category assigned by the IUCN, the common habitat, and a detailed description of the species. In the top right-hand corner, there is a button that allows the user to delete this reading. If this button is clicked, once the reading has been removed, the message "Reading successfully deleted" is displayed, as shown in Figure 21. This user-initiated deletion process simplifies data management and improves the application's functionality.

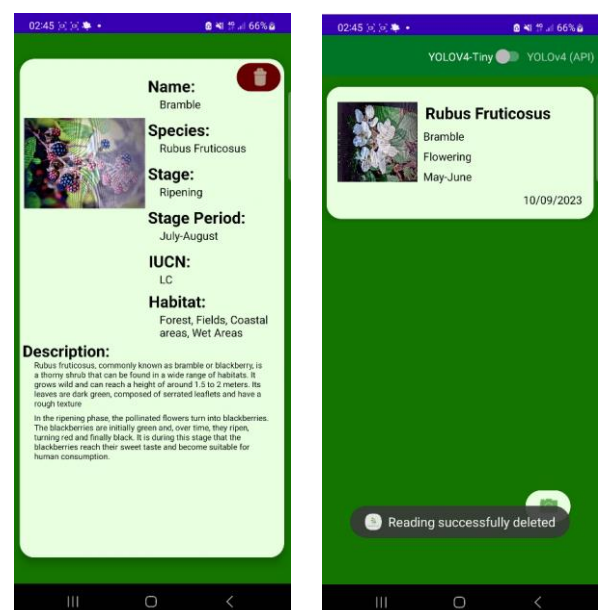


Figure 18: The interfaces showing the details and the message with the deletion notification.

In the top corner of the main menu (Figure 20), the user can choose the CNN model they want to use to detect and classify plants: YOLOv4 or YOLOv4-tiny. In the bottom right-hand corner, there is a button that activates the mobile device's camera to take a photo. This image is then transmitted to the model selected for analysis. Figure 22 shows the process of capturing an image (i.e., photo) to be analyzed.

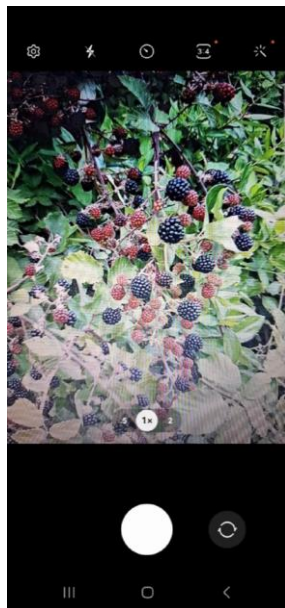


Figure 19: The process of capturing an image.

If the user has selected the YOLOv4 model and there is no Internet connectivity to the API, he will receive the message "No connection to the API, use the local model", shown in Figure 23. This information allows the user to decide whether to use YOLOv4-tiny for local processing in the main menu (Figure 20).

In cases where the chosen model cannot detect the plant and its development stage, the user will receive the message "Bad reading, please try again", shown in Figure 23. These error messages are enlightening in that they suggest possible causes to help the user.

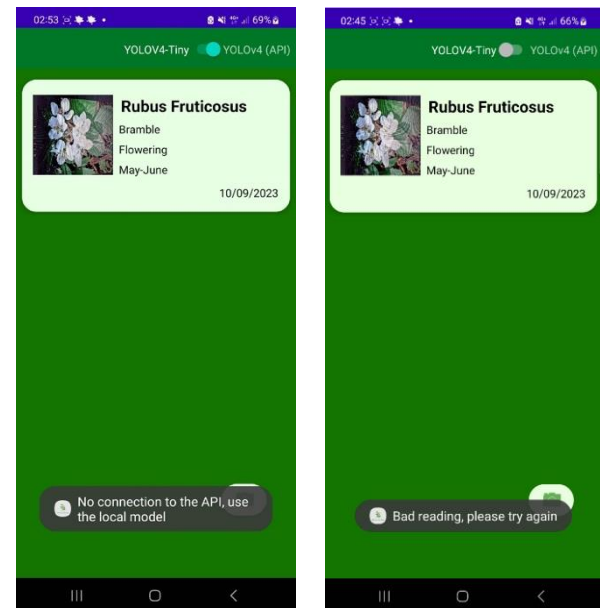


Figure 20: The error messages notification.

3 Conclusion

Wild flowers and plants are a vital part of biological diversity and an essential resource for the planet. However, we are seeing an increase in the number of wild flowers and plants at risk of extinction and in decline due to climate change and the impact of human action. Therefore, there is an urgent need to contribute technological solutions for their conservation and preservation.

The work presented in this article is one of the stages of an ongoing research project, which aims to develop a mobile application and a system based on computer vision techniques to detect and monitor the development stages of wild flowers and plants.

In summary, the main contributions resulting from this article are: 1) the creation of a dataset with the stages of development of a wild plant; 2) a comparative performance analysis of the YOLOv4 and YOLOv4-tiny convolutional neural network models for detecting the development stages of this wild plant; 3) a description of the process of developing a mobile application, using YOLOv4 and YOLOv4-tiny, as a proof of concept.

This mobile application can be used by visitors to a nature park to provide information and raise awareness about the development stages of the wild flowers and plants they encounter along the roads and trails. The application is currently in the testing phase of its prototype version.

Several points remain open for future work, including: 1) creating a dataset with support for a wide range of wild flowers and plants species, with a large number of images for each plant species and development stage; 2) testing and evaluating other convolutional neural network models; 3) continuing the process of validating the application with a wide range of real users; 4) using the feedback from these users to add features to the application that enrich and facilitate its use.

Acknowledgments

J.M.L.P.C. and V.N.G.J.S. acknowledge that this work is funded by FCT/MCTES through national funds and when applicable co-funded EU funds under the project UIDB/50008/2020. P.D.G. thanks the support provided by the Center for Mechanical and Aerospace Science and Technologies (C-MAST) under project UIDB/00151/2020.

This is within the activities of project Montanha Viva – An intelligent prediction system for decision support in sustainability, project PD21-00009, promoted by PROMOVE program funded by Fundação La Caixa and supported by Fundação para a Ciência e a Tecnologia and BPI.

Declarations

Author contributions. Conceptualization, P.D.G., J.V; methodology, J.V; validation, P.D.G., J.M.L.P.C. and V.N.G.J.S.; formal analysis, P.D.G., J.M.L.P.C. and V.N.G.J.S.; investigation, J.V; writing—original draft preparation, J.V; writing—review and editing, P.D.G., J.M.L.P.C. and V.N.G.J.S.; supervision, J.M.L.P.C. and V.N.G.J.S.; funding acquisition, P.D.G., J.M.L.P.C. and V.N.G.J.S. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest. The authors declare no conflict of interest.

References

- [1] Agence France-Presse, “Chain-reaction extinctions will cascade through nature: Study | Daily Sabah.” <https://www.dailysabah.com/life/environment/chain-reaction-extinctions-will-cascade-through-nature-study> (accessed Jan. 29, 2023).
- [2] L. E. Grivetti and B. M. Ogle, “Value of traditional foods in meeting macro- and micronutrient needs: the wild plant connection,” *Nutr Res Rev*, vol. 13, no. 1, pp. 31–46, Jun. 2000, doi: 10.1079/095442200108728990.
- [3] E. Christaki and P. Florou-Paneri, “Aloe vera: A plant for many uses,” *J Food Agric Environ*, vol. 8, pp. 245–249, 2010.
- [4] Trevor Dines, “Plantlife - A Voice for Wildflowers - Ark Wildlife UK.” <https://www.arkwildlife.co.uk/blog/plantlife-a-voice-for-wildflowers/> (accessed Jan. 29, 2023).
- [5] X. Chi *et al.*, “Threatened medicinal plants in China: Distributions and conservation priorities,” *Biol Conserv*, vol. 210, Part A, pp. 89–95, Jun. 2017, doi: 10.1016/J.BIOCON.2017.04.015.
- [6] Woodstream, “Learn The Six Plant Growth Stages.” <https://www.saferbrand.com/articles/plant-growth-stages> (accessed Sep. 04, 2023).
- [7] João Videira, Pedro D. Gaspar, Vasco N. G. J. Soares, and João M. L. P. Caldeira, “Detecting and Monitoring the Development Stages of Wild Flowers and Plants using Computer Vision: Approaches, Challenges and Opportunities (in press),” *International Journal of Advances in Intelligent Informatics (IJAIN)*, 2023.
- [8] PEAT GmbH, “Plantix - seu médico agrícola – Apps no Google Play.” https://play.google.com/store/apps/details?id=com.peat.GartenBank&hl=pt_PT&gl=US (accessed Oct. 06, 2022).
- [9] AIBY Inc., “Plantum - Identificar plantas – Apps no Google Play.” https://play.google.com/store/apps/details?id=plant.identification.flower.tree.leaf.identifier.identify.cat.dog.breed.nature&hl=pt_PT&gl=US (accessed Oct. 06, 2022).
- [10] N. Buch, S. A. Velastin, and J. Orwell, “A review of computer vision techniques for the analysis of urban traffic,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 3, pp. 920–939, Sep. 2011, doi: 10.1109/TITS.2011.2119372.
- [11] S. Xu, J. Wang, W. Shou, T. Ngo, A. M. Sadick, and X. Wang, “Computer Vision Techniques in Construction: A Critical Review,” *Archives of Computational Methods in Engineering* 2020 28:5, vol. 28, no. 5, pp. 3383–3397, Oct. 2020, doi: 10.1007/S11831-020-09504-3.
- [12] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan, “Contextualizing object detection and classification,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37., 20015, pp. 13–27. doi: 10.1109/CVPR.2011.5995330.
- [13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [14] The MathWorks Inc., “What Is Object Detection? - MATLAB & Simulink.” https://www.mathworks.com/discovery/object-detection.html?s_tid=srchtitle_object%20detection_1 (accessed Dec. 26, 2022).
- [15] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, “YOLOv4: Optimal Speed and Accuracy of Object Detection.” arXiv, 2020. doi: 10.48550/ARXIV.2004.10934.
- [16] G. Li, X. Huang, J. Ai, Z. Yi, and W. Xie, “Lemon-YOLO: An efficient object detection method for lemons in the natural environment,” *IET Image*

- Process*, vol. 15, no. 9, pp. 1998–2009, Mar. 2021, doi: 10.1049/ipr2.12171.
- [17] A. Shill and M. A. Rahman, “Plant disease detection based on YOLOv3 and YOLOv4,” *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0, ACMI 2021*, pp. 1–6, Jul. 2021, doi: 10.1109/ACMI53878.2021.9528179.
- [18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, doi: 10.1109/cvpr.2016.91.
- [19] J. Redmon and A. Farhadi, “YOLOv3: An Incremental Improvement,” Apr. 2018, Accessed: Aug. 21, 2023. [Online]. Available: <https://arxiv.org/abs/1804.02767v1>
- [20] The MathWorks Inc., “Anchor Boxes for Object Detection - MATLAB & Simulink.” <https://www.mathworks.com/help/vision/ug/anchor-boxes-for-object-detection.html> (accessed Dec. 26, 2022).
- [21] Q. Chen and Q. Xiong, “Garbage Classification Detection Based on Improved YOLOV4,” *Journal of Computer and Communications*, vol. 8, pp. 285–294, 2020, doi: 10.4236/jcc.2020.812023.
- [22] Z. Jiang, L. Zhao, S. Li, Y. Jia, and Z. Liqun, “Real-time object detection method based on improved YOLOv4-tiny,” *Journal of Network Intelligence*, vol. 7, no. 1, Nov. 2022, Accessed: Aug. 23, 2023. [Online]. Available: <https://arxiv.org/abs/2011.04244v2>
- [23] L. Song *et al.*, “Object detection based on Yolov4-Tiny and Improved Bidirectional feature pyramid network,” *Journal of Physics: Conference Series* *2021 International Conference on Electronic Communication, Computer Science and Technology 07/01/2022-09/01/2022 Nanchang*, vol. 2209, no. 1, Feb. 2022, doi: 10.1088/1742-6596/2209/1/012023.
- [24] W. Zhang *et al.*, “Airborne infrared aircraft target detection algorithm based on YOLOv4-tiny,” *Journal of Physics: Conference Series* *2021 International Conference on Advances in Optics and Computational Sciences (ICAOS) 2021 21-23 January 2021, Ottawa, Canada*, vol. 1865, no. 4, Apr. 2021, doi: 10.1088/1742-6596/1865/4/042007.
- [25] Ken-ichi Ueda, Nate Agrin, and Jessica Kline, “Uma comunidade para naturalistas · iNaturalist.” <https://www.inaturalist.org/> (accessed Aug. 21, 2023).
- [26] Ken-ichi Ueda, Nate Agrin, and Jessica Kline, “Uma comunidade para naturalistas · BioDiversity4All.” <https://www.biodiversity4all.org/> (accessed Aug. 23, 2023).
- [27] João Videira, Pedro D. Gaspar, Vasco N. G. J. Soares, and João M. L. P. Caldeira, “MontanhaVivaApp Dataset | Kaggle.” <https://www.kaggle.com/datasets/krosskrosis/montanhavivaapp-dataset> (accessed Sep. 05, 2023).
- [28] Yonghye Kwon, “GitHub - developer0hye/Yolo_Label: GUI for marking bounded boxes of objects in images for training neural network YOLO.” https://github.com/developer0hye/Yolo_Label (accessed Aug. 21, 2023).
- [29] Alexey Bochkovskiy, “GitHub - AlexeyAB/darknet: YOLOv4 / Scaled-YOLOv4 / YOLO - Neural Networks for Object Detection (Windows and Linux version of Darknet).” <https://github.com/AlexeyAB/darknet> (accessed Aug. 21, 2023).
- [30] J. A. Cook and J. Ranstam, “Overfitting,” *British Journal of Surgery*, vol. 103, no. 13, p. 1814, Dec. 2016, doi: 10.1002/bjs.10244.
- [31] M. Decuyper, M. Stockhoff, S. Vandenbergh, al -, and X. Ying, “An Overview of Overfitting and its Solutions,” *J Phys Conf Ser*, vol. 1168, no. 2, p. 022022, Feb. 2019, doi: 10.1088/1742-6596/1168/2/022022.
- [32] L. Prechelt, “Early Stopping - But When?,” pp. 55–69, 1998, doi: 10.1007/3-540-49430-8_3.
- [33] The Interaction Design Foundation, “What is User Centered Design?” <https://www.interaction-design.org/literature/topics/user-centered-design> (accessed Sep. 08, 2023).
- [34] Eastern Peak, “Iterative Development.” <https://easternpeak.com/definition/iterative-development/> (accessed Sep. 08, 2023).
- [35] Matt Rae, “What is Adobe XD and What is it Used for?” <https://www.adobe.com/products/xd/learn/get-started/what-is-adobe-xd-used-for.html> (accessed Sep. 08, 2023).
- [36] Pavel Gorbachenko, “Functional vs Non-Functional Requirements | Enkonix.” <https://enkonix.com/blog/functional-requirements-vs-non-functional/> (accessed Sep. 08, 2023).
- [37] IBM, “Use-case diagrams - IBM Documentation.” <https://www.ibm.com/docs/en/rational-software/9.6.1?topic=diagrams-use-case> (accessed Sep. 08, 2023).
- [38] Google and JetBrains, “Android Studio & App Tools - Android Developers.” <https://developer.android.com/studio> (accessed Sep. 08, 2023).
- [39] D. Richard Hipp, “SQLite.” <https://www.sqlite.org/index.html> (accessed Sep. 08, 2023).
- [40] Armin Ronacher, “Welcome to Flask.” <https://flask.palletsprojects.com/en/2.3.x/> (accessed Sep. 08, 2023).
- [41] João Videira, Pedro D. Gaspar, Vasco N. G. J. Soares, and João M. L. P. Caldeira, “MontanhaVivaApp – Google Drive.” <https://drive.google.com/drive/u/2/folders/1FX6pwvDgV2lN9u3EwtH66DhPunIPJ3AT> (accessed Aug. 24, 2023).
- [42] Việt Hùng, “GitHub - hunglc007/tensorflow-yolov4-tflite: YOLOv4, YOLOv4-tiny, YOLOv3, YOLOv3-tiny Implemented in Tensorflow 2.0, Android. Convert YOLO v4 .weights tensorflow, tensorrt and tflite.” <https://github.com/hunglc007/tensorflow-yolov4-tflite> (accessed Aug. 21, 2023).

- [43] João Videira, Pedro D. Gaspar, Vasco N. G. J. Soares, and João M. L. P. Caldeira, “videira202011/MontanhaVivaApp.” <https://github.com/videira202011/MontanhaVivaApp> (accessed Sep. 04, 2023).
- [44] Oracle, “Creating and Using Packages (The Java™ Tutorials > Learning the Java Language > Packages).” <https://docs.oracle.com/javase/tutorial/java/package/packages.html> (accessed Sep. 05, 2023).

Image Segmentation: A Modern Roadmap

Matilde Delgado de Sousa^{1,2*}, Pedro Dinis Gaspar^{1,2} and
Nuno Pereira^{1,2,3}

^{1*}Department of Electromechanical Engineering, Faculty of Engineering,
University of Beira Interior, Rua Marquês d'Ávila e Bolama, Covilhã,
6201-001, Castelo Branco, Portugal.

²Department of Computer Engineering, Faculty of Engineering,
University of Beira Interior, Rua Marquês d'Ávila e Bolama, Covilhã,
6201-001, Castelo Branco, Portugal.

³LITecS - Laboratory of Innovation and Technologies for Sustainability, C-
MAST—Center for Mechanical and Aerospace Science and Technologies,
Calçada Fonte do Lameiro 6, Covilhã, 6200-358, Castelo Branco, Portugal.

*Corresponding author(s). E-mail(s): galvao.sousa@ubi.pt;
Contributing authors: dinis@ubi.pt; nuno.pereira@ubi.pt;

Abstract

Image segmentation is a fundamental task in computer vision that has rapidly evolved through advanced models and methodologies. Keeping pace with these developments can be challenging, particularly for researchers aiming to identify the most appropriate techniques for their specific applications. This roadmap is designed to provide researchers with the necessary tools and insights to explore the complex landscape of image segmentation and select the most suitable models for their tasks. This work presents a comprehensive roadmap that explores the key concepts, 17 datasets, and evaluation metrics. It includes a review of a total of 111 models and methodologies organized accordingly. To complement the roadmap, an interactive tool was developed to visualize model interconnections, compare performance metrics such as speed, average precision, panoptic quality and mean intersection union across different tasks, and directly access the original research. The objective of this work is to provide researchers with a comprehensive and accessible resource that supports an effective selection and application of image segmentation models, to continue advancing in the field of image segmentation and computer vision.

Keywords: Computer Vision, Image Segmentation, Instance Segmentation, Panoptic Segmentation, Deep Learning, Interactive Interface

1 Introduction

Nowadays, computer vision is commonly seen in our daily lives, from face recognition in our phones to object classification using a single image, from computer designing images to autonomous cars. However, this emerging technology took years to reach this state. In the 60's Papert, S.A. launched "The Summer Vision Project" which aimed to develop a system to perform object identification, by teaching computers, patterns and correlations between objects [1]. Since then, the computer vision world with the aid of the constant technological evolution has developed multiple new solutions and new algorithms that are more precise and essentially fast.

Computer vision is a subfield of Artificial Intelligence. Computer vision allows systems and computers to exploit information from images and videos to perform a task using images, video and cameras. This paper explores object segmentation, identification, and localization, which are just some of the many practical applications of computer vision. There are several methods to accomplish object identification, but the most popular is object segmentation or image segmentation. Image segmentation aims to classify each pixel present in the image according to its visual characteristics, before grouping them into distinct clusters based on their visual similarity. A cluster of pixels belonging to the same class is called a segment. With a segmented image, pixels can be thought of as class labels rather than real pixel values [2]. It is useful to divide objects into two categories: Stuff and Things. Things refer to objects that are properly bounded and can be counted, such as humans. Stuff is all of the things that are not geometrically defined but are identified by the material, such as the sky [3].

Semantic segmentation, instance segmentation, and panoptic segmentation are specialized forms of image segmentation in computer vision. In semantic segmentation, each pixel in an image is categorized into a specific class, but individual instances of the same class are not distinguished. Instance segmentation extends this by not only classifying each pixel into a category but also identifying separate instances of objects within the same class. Panoptic segmentation unifies these approaches, offering both class labels and individual instance identifiers for each pixel in an image. The key differences among these segmentation methods lie in their granularity and the level of detail they provide: Semantic segmentation is geared towards general class-level categorization; instance segmentation offers additional instance-level identification within each class; and panoptic segmentation provides a comprehensive, unified labeling, effectively combining the features of both semantic and instance segmentation. The distinction between these three segmentation methods can be seen in Figure 1:

Additionally, there are two classes of segmentation raising popularity, which each of the methods falls, interactive segmentation and automatic segmentation. In interactive segmentation, a person is required to refine the resulting mask of any class of object by clicking on it, to guide the deep neural network, where the input is used as supervised information [4]. Automatic segmentation is a fully automated process that relies only on algorithms to do the partition, thus categorizing specific objects ahead of time. Nevertheless, it requires a substantial amount of fully annotated data to train the segmentation model.

[5] introduced the One-shoot Learning approach, in which the model uses one example (or a very small number of them) and predicts the class of an unseen object by

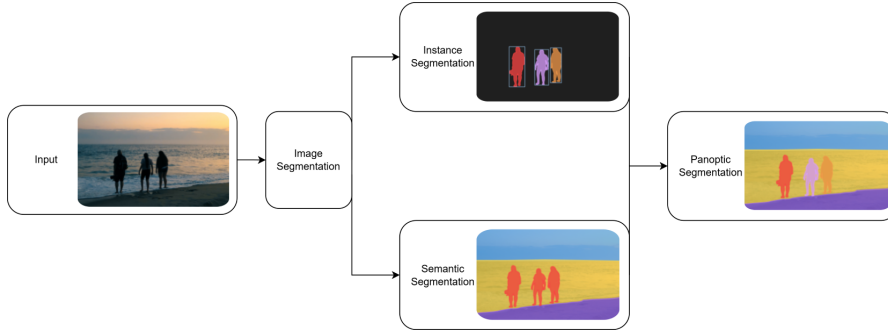


Fig. 1 Representation of three different methods of segmentation. (*Up*) Instance Segmentation: attributes unique labels to each object. (*Down*) Semantic Segmentation: labels equally the same object class. (*Right*) Panoptic Segmentation: recognizes instances while at the same time recognizing the items according to their class

analyzing the similarities of both data. This method addresses the need for enormous amounts of annotated data for a segmentation problem. This idea can be modified by the few-shot learning method, which uses a small sample of data to achieve similar results.

Afterward, [6] introduced the concept of Zero-data Learning, which would later be commonly known as Zero-Shoot Learning or ZSL. The assumption is that the model can generalize classes or tasks with no training data and just a description of the classes. Consequently, the model would be able to segment an object using just prompt text or an image. Nevertheless, this method has a tendency to favor seen classes [7]. In other words, these methods often don't have instances of unseen objects and instead consider them as background during the training phase. This limits the model's exploration of the relationships between unseen classes and causes it to lose valuable visual information. Consequently, during prediction, the model distinguishes new classes purely based on their predefined semantic representations [8]. To tackle this issue, [9] and [10] used the concept of open-vocabulary image segmentation. Open-vocabulary image segmentation is a technique designed to divide an image into distinct semantic regions based on text descriptions, which can be entirely arbitrary and not limited to pre-defined categories. Unlike traditional image segmentation methods that rely on a fixed set of labelled categories, open-vocabulary image segmentation allows for more flexibility and adaptability by interpreting and segmenting images according to any text input provided. This means the system can recognize and limit various objects and regions within an image, even if those objects and regions were not part of its original training dataset. By leveraging advanced models that align visual elements with linguistic descriptions, open-vocabulary image segmentation can organize pixels into meaningful groups that correspond to the specified text descriptions, enabling a more dynamic and comprehensive understanding of the visual content.

This roadmap is structured into six main sections: Annotations, Datasets and Metrics, Image Segmentation Methods, Exploration Framework, Challenges and Future Prospects, and Conclusion. The Annotations section highlights the significance of

annotations in a computer vision testing pipeline. It is further divided into six subsections, each introducing a current annotation method. The Datasets and Metrics section discusses the usefulness of datasets and provides examples of datasets employed in computer vision. Additionally, it offers an overview of various metrics employed when testing models. The Image Segmentation section consists of four subsections presenting a brief explanation of key concepts of these methods, the description of some common backbones used for these state-of-the-art methods and the state-of-the-art approaches to semantic, instance, and panoptic segmentation. The Exploration Framework ¹ is an interactive tool built upon the literature review presented in this paper. Its purpose is to help researchers visualize the connections between various papers, explore the most effective metrics and models for each task, and provide an innovative way to navigate through the literature review produced in this paper. The next section discusses future challenges and prospects for the field of image segmentation. Finally, the Conclusion section summarizes the roadmap.

The contributions of this work are summarized as follows:

- Annotations used in image segmentation datasets are explained and compared in terms of accuracy and applicability.
- Seventeen common benchmark datasets used to assess image segmentation models are divided into objectives, and then explained and contrasted with various parameters.
- 5 evaluation metrics are defined, and numerous comparisons of the most significant works in the state-of-the-art are made to demonstrate their performance across different datasets and criteria.
- Following the presentation of the background of image segmentation technology and its key features, a full taxonomy of 111 existing works on the three main areas, such as the methodology used to design image segmentation models and application scenarios, is carried out.
- The development of an interactive exploration framework (<https://github.com/Matilde3Sousa/Roadmap>) to help researchers find the appropriate model for their intended tasks.
- Following a discussion of current challenges, the potential opportunities for the future are presented.

2 Annotations

Annotations are the process of manually labeling and tagging specific pixels or regions in an image to distinguish and identify objects or regions of interest. When training a model, these annotations are the ground truth against which predictions are compared. Therefore, annotations play an important role in testing effective image segmentation models. There are three main tasks in computer vision where the annotations are essential: image classification, object detection, and image segmentation.

¹The exploration framework designed for this roadmap is available online at <https://github.com/Matilde3Sousa/Roadmap>.

Image classification has a fairly simple annotation technique, i.e., it requires only the captioning of an image. These labels can be simple text, class numbers, or one-hot encode tensors that provide a particular ID. Object detection is more difficult, yet simpler than image segmentation since object detection requires only the location of the object and its class. Therefore, bounding boxes are usually used, which are then considered as ground truth. Lastly, annotation is more challenging in image segmentation since the goal is to have the most precise outline of the object in order to obtain a model with greater precision. Usually, segmented masks, which include polygonal masks, or binary masks are used for this purpose, which represent the boundaries of the object or scene and are marked with a specific class.

2.1 Bounding Boxes

Bounding boxes are arguably one of the simplest forms of annotation, second only to image classification. Essentially, a bounding box is a rectangular outline that encloses an object within an image. The borders of the box delineate the spatial extent of the object, which is subsequently associated with a specific class label. Though simpler and less precise than other annotation methods, bounding boxes are predominantly employed in object detection tasks.

2.2 Mask

In mask annotations, each pixel in a designated area is shaded to indicate key features and is labelled as either a region of interest or as a background. Unlike bounding boxes, which typically use rectangular shapes, mask annotations offer more granularity and can be complex in shape.

Various algorithms can generate these masks, which may either be binary—representing a single class—or multi-colored to distinguish multiple classes. A specialized form of mask annotation used in object detection and semantic segmentation is the polygonal mask. Polygonal masks are more precise due to their multiple vertices, providing greater flexibility in the annotation process. Furthermore, polygonal masks are efficient in terms of storage space and can be readily converted into vector format, offering an optimal balance between accuracy and computational efficiency.

While binary and polygonal masks offer fundamental representations, more advanced techniques, namely semantic and instance segmentation, have emerged as refined means of creating masks. Semantic segmentation annotation is a process that involves assigning a specific class label to each pixel in an image based on the semantic meaning of the object or region to which it belongs. This technique divides the image into different segments, depending on the structures or objects present in it. The output of this annotation is a high-resolution map that predicts the class of each pixel. Semantic segmentation proves especially valuable for tasks that demand precise location or boundary information about objects or regions in an image. On the other hand, the process of instance segmentation annotations involves assigning a distinct label to each pixel in an image, corresponding to a particular object instance. This technique comes in handy when there is a requirement to differentiate between

multiple instances of the same category. Unlike semantic segmentation annotation, which groups pixels into classes, this approach ensures that each object instance in the image is labelled distinctly. It proves to be particularly useful when there is a need to accurately locate an object or differentiate within the same semantic class.

3 Datasets and Metrics

When developing solutions in computer vision, evaluating performance is paramount for ensuring real-world applicability. The evaluation tools are predominantly categorized as databases or datasets. These datasets serve as an organized compendium of images, meticulously curated for specialized objectives such as classification, segmentation, or object recognition. Each dataset encompasses a set of data instances or samples, each endowed with an array of features aimed at assisting the computer vision algorithm.

In addition to datasets, another indispensable element for evaluation is metrics. Metrics are formulated sets of rules and mathematical equations employed to gauge the outcomes of different computer vision methods. Metrics may include, but are not limited to, accuracy, precision, recall, F1-score, and Intersection over Union (IoU) for various tasks such as classification and object detection. These metrics offer quantitative insights that aid in the iterative process of model refinement. The metrics will be discussed later in the manuscript.

3.1 Datasets

The following section introduces numerous datasets with various characteristics, such as the types of scenes, annotations, and the overall quantity of images. A review of this section can be seen in Table 1.

3.1.1 Cityscapes

The Cityscapes dataset was created to fill the gap of understanding urban street scenes, essential for autonomous driving. [11] acquired data for several months in multiple cities during the daytime, ultimately gathering over 25,000 images ready to be used in segmentation training. The dataset contains 30 classes created by polygonal annotations, which includes 5,000 fine-annotated images and 20,000 coarse-annotated images. These classes go from people to the road to the sky to vehicle. Furthermore, other researchers have added some extensions to the dataset such as bounding box annotations and augmented data by adding fog and rain. Cityscapes presents four benchmark challenges, namely pixel-level, instance-level, panoptic semantic labeling, and 3D vehicle detection. These tasks are benchmarks for multiple authors to rank their models.

3.1.2 COCO

In urban scene recognition, [12] developed a dataset containing objects from everyday life in their natural status. The Common Objects in Context, or COCO, contains 330,000 images with over 91 object classes and with per-instance-segmentation that

provides exact localization and labeling of objects. This dataset is one of the most used datasets for object detection and segmentation. Additionally, it was created a ramification of this dataset named Common Objects in Context-stuff, or COCO-stuff [3] which focuses on stuff classes. Including the original dataset semantic segmentation annotations for scene understanding tasks, augmenting to 172 thing and stuff classes. This upgrade enables researchers to do scene parsing. In addition to the main dataset, there are several subsets available for detection, panoptic and keypoints tasks. Participants can submit their work for each of these subdatasets and get ranked among other works. The latest update for these tasks was in 2020. Despite this, COCO tasks are still highly popular in the research community as a means of ranking models.

3.1.3 Pascal VOC

Pascal Visual Object Classes is a dataset aimed to evaluate classification, object detection, image segmentation, and people layout models created by [13]. It contains about 23,060 images divided into 20 classes such as person, chair, car, and sky. The sum of the training set and the validation set reaches 11,530 with 27,450 objects annotated by the region of interest and with about 6,929 segmentations. The VOC 2012 challenge includes six different tasks: Classification, Detection, Segmentation, Action Classification, ImageNet Large Scale Visual Recognition, and Person Layout. The first two challenges are straightforward: classifying and detecting a class. The Segmentation task generates pixel-wise segmentations, while the Action Classification involves predicting an action. The ImageNet Large Scale Visual Recognition task is for automatic annotation, and the Person Layout task is for predicting the body parts of a person. This dataset was regularly updated until 2012 and is still one of the benchmark challenges for computer vision models.

3.1.4 KITTI

KITTI, or Karlsruhe Institute of Technology and Toyota Technological Institute, is a popular dataset used on autonomous driving. [14] gathered hours of traffic data of the city of Karlsruhe, Germany, through multiple devices such as grayscale stereo cameras and high-resolution RGB that were append on a driving platform. Nevertheless, the dataset does not provide instance segmentation annotations. Annotations like bounding boxes or instance segmentation were developed lately by other researchers to be applied to their projects. It has a total of 34 classes identified. KITTI has multiple sub-datasets for different tasks such as stereo, optical flow, sceneflow, object detection, semantic/instance, etc. Each task has an evaluation metric and respective evaluation website, where is showcased the ranked models.

3.1.5 CamVid

CamVid [15, 16] is a dataset designed for semantic segmentation tasks in the context of scene understanding and autonomous driving. It provides real-world urban driving scenarios that allow researchers to evaluate the performance of their models under such conditions. The dataset comprises over 700 road images captured from a camera mounted on a vehicle, which provides a perspective similar to that of a driver.

The images were captured from five video sequences and included various conditions such as traffic, different infrastructures, vehicles, roads, and nature. Each frame has a resolution of 960x720 and was manually annotated with semantic masks providing 32 semantic classes such as bicyclist, train, void, tree, and road.

3.1.6 Mapillary Vistas dataset (MVD)

The Mapillary Vistas Dataset (MVD) [17] is a benchmark dataset that focuses on street-level images. It contains over 25,000 images captured from multiple devices, such as mobile phones and action cameras, across six continents and various weather conditions. Captured by multiple devices ranging from mobile phones to action cameras, this dataset offers a diverse range of urban environments, objects, road conditions, and landmarks, each with multiple image qualities. MVD comes in two versions: v1.2 with 66 object categories, 30 of which are pixel-level labels, and v2.0 with 124 semantic object classes and 70 instance-specifically labelled categories. Polygonal annotation is used throughout the dataset. The primary metrics for evaluating models on MVD are mean intersection over union (mIoU) for semantic tasks and average precision (AP) for instance tasks. Furthermore, Mapillary has five additional datasets tailored for specific tasks: Mapillary CrowdDriven Dataset, Mapillary Metropolis Dataset, Mapillary Planet-Scale Depth Dataset, Mapillary Street-level Sequences Dataset, and Mapillary Traffic Sign Dataset.

3.1.7 Grand Theft Auto 5 Synthetic Images

The Grand Theft Auto 5 Synthetic Images dataset, as described in [18], is a collection of synthetic images that were generated from video games to train computer vision models. The hyper-realistic world of Grand Theft Auto makes it an ideal source for synthetic images. The dataset contains 25,000 images, each frame measuring 1914×1052 pixels, extracted from the game. These images were meticulously annotated using specialized software, completing the process in just 49 hours, a remarkable feat compared to annotation processes in other datasets. The annotations cover 19 semantic classes, including car, road, and sky. The study revealed that models trained on a combination of synthetic and real-world data outperformed those trained with only real-world data. This finding could be a solution to the challenge of needing vast amounts of data for training computer vision models.

3.1.8 Large Vocabulary Instance Segmentation (LVIS)

The Large Vocabulary Instance Segmentation (LVIS) [19] is a benchmark dataset aiming to help improve instance segmentation. The dataset is an extension of COCO and is used annually on challenges that require object detection and instance segmentation. It provides 164,000 images from common objects and everyday scenes. Up to now, LVIS contains 1,200 object categories, some rare with only a few images, with high-quality annotations masks, and over 2.2 million instances. It is divided into a training dataset with 52,263 images and 693,958 instances and a validation dataset with 5,000 images with over 50,763 instances.

3.1.9 KITTI INStance dataset (KINS)

Qi and his team, as detailed in their research paper [20], expanded the KITTI dataset by providing instance and semantic annotations for nearly 15,000 images. These annotations include amodal instance masks, relative occlusion order, and semantic labels. The KITTI Instance dataset serves as a comprehensive amodal instance dataset meant for autonomous driving. It can not only aid in instance segmentation but also amodal instance segmentation that can extend to scene flow estimation. The dataset comprises 7,474 training images and 7,517 testing images and is divided into two main categories: vehicles and people. The vehicles category includes classes like trucks and cars, while the people category accounts for pedestrians and cyclists. The dataset’s primary objective is to assist with autonomous driving and has thus been segregated into these two categories.

3.1.10 Open Image

The Open Image dataset [21], which was first released in 2016, has now reached its seventh version and boasts 9 million annotated images. This dataset comprises images of everyday objects and urban scenes. It provides several types of annotations, including 1.9 million images with bounding boxes for over 600 object classes, 3.3 million annotations with 1,466 visual relationships, 2.8 million object instances with over 350 classes, 1.4 million images with 66.4 million point-level labels covering 5,827 classes, and 61.4 million image-level labels generating 20,638 classes. The dataset also organizes classes into a hierarchical structure, providing researchers with more detailed and specific knowledge about the different elements and classes. In addition, Version 6 of the Open Image dataset includes 675,000 localized narratives, making it the largest dataset available with object location annotations. The dataset also has three extensions: HierText, which is a hierarchical annotated dataset of text in natural scenes and documents aimed at helping Optical Character Recognition (OCR) researchers develop more robust OCR models; MIAP, or More Inclusive Annotations for People, which includes an additional 100,000 fully annotated images with specific features of people to broaden the understanding and inclusivity of models; and Crowdsourced, which comprises over 382,000 images in 6,000 classes developed by users of the Google Crowdsourcing App.

3.1.11 Open Plant Phenotyping Database (OPPD)

The Open Plant Phenotyping Database (OPPD) [22] is a dataset that contains images of different types of plant seedlings. It is useful for problems related to visual recognition. The dataset includes 7,590 images of 47 plant species that were grown under three different conditions: ideal, drought, and natural. The plants were tracked and photographed at various stages of growth, including sowing. The images have an average resolution of 44 pixels per square millimeter. The dataset includes bounding box annotations and can be used for tasks such as instance detection and plant species classification.

3.1.12 ADE20K

The ADE20K dataset is a specialized collection of images used for object detection and image segmentation. It was developed by [23] and [24] and consists of 27,000 images taken from the Sun and Places database. The dataset contains almost 3,000 annotated object categories, offering pixel-level labels for a wide range of objects and stuff classes. It includes images of everyday scenes, featuring people, animals, vehicles, and more, with some pictures even including object components and parts. Additionally, the dataset is designed to protect privacy by masking faces and license plates. The ADE20K dataset has been used as a benchmark in multiple computer vision challenges, especially in segmentation.

3.1.13 SA-1B

The SA-1B dataset, created by [25], is part of the Segment Anything project, which aims to establish a foundational model for image segmentation. Its purpose is to aid researchers in training and evaluating models. To achieve this, the dataset includes 11 million high-resolution images from photographers worldwide and 1.1 billion mask segmentations collected by the Segment Anything engine. On average, each image has 100 masks and a resolution of 1500x2500 pixels. The SA-1B dataset is class-agnostic, which means that the masks are not specific to any class. However, the dataset ensures that faces and license plates are de-identified.

3.1.14 Indian Driving Dataset (IDD)

The Indian Driving Dataset (IDD) [26] provides a challenging dataset to improve the autonomous driving task. By leveraging the traffic behavior and road conditions specific to India, it offers a more complex and unstructured environment for models to train on. The dataset included images captured from highways, as well as urban and rural settings. It consists of 10,004 finely annotated labelled images organized into 34 classes. The authors developed a four-level labeling hierarchy to address labeling ambiguity and to provide researchers with options regarding the level of detail needed for their models.

3.1.15 Berkeley Deep Drive (BDD100K)

[27] created the BDD100K dataset to tackle the challenges in autonomous driving. This dataset consists of over 100,000 videos, each lasting 40 seconds, recorded under various weather conditions, times of the day, and driving scenarios. It can be utilized for multiple tasks such as: Image Tagging, Object Detection, Lane Marking, Drivable Area, Multiple Object Tracking, Imitation Learning and both Instance and Semantic Segmentation. For instance and semantic segmentation, the dataset contains 40 classes with 10,000 randomly sampled frames, with masks for each instance as well as pixel-level annotations for semantic segmentation. The dataset is split into training (7,000 images), validation (1,000 images) and test (2,000 images) sets.

3.1.16 Adverse Conditions Dataset (ACDC)

The Adverse Conditions Dataset [28], or ACDC, is designed to improve semantic segmentation in driving scenarios and supports uncertainty-aware semantic segmentation. The authors had gathered 4006 images of road under adverse conditions — fog, snow, nighttime, and rain— with an equal distribution among these four categories. Each image features high-quality, fine pixel-level semantic annotations, includes a corresponding image of the same scene under normal conditions, and provides a binary mask that distinguishes regions with clear (valid) semantic content from those that are ambiguous (invalid). The dataset includes 19 classes, matching both in number and classification to those in the Cityscapes [11] dataset.

3.1.17 IDD-AW

Following its precursor IDD [26], [29] takes advantage of the unstructured environment of India to provide a more challenging and diverse dataset, called IDD-AW. The authors generate a new improved dataset by capturing images in adverse weather conditions and gathering spectral data (near-infrared imaging), which aims to improve segmentation accuracy and safety. The dataset contains 5,000 images distributed across four adverse conditions: low light (1,000), foggy (1,500), snowy (1,000) and rainy (1,500). Each image is annotated with dense pixel-level semantics. Furthermore, the authors implement a four-level hierarchy label system covering 30 classes, helping to reduce ambiguity and allowing for varying levels of granularity in segmentation.

3.2 Metrics

This section provides a detailed analysis of five critical metrics for the evaluation of the model performance in the image segmentation field. These metrics include Mean Intersection over Union (mIoU) and Pixel Accuracy (PA), both of which are predominantly utilized in the evaluation of semantic segmentation models. For assessing performance in instance segmentation, Average Precision (AP) serves as the main metric. Additionally, Panoptic Quality (PQ) is employed as a metric for gauging the effectiveness of models in panoptic segmentation. Furthermore, Frames Per Second (FPS) are introduced, a metric that is increasingly gaining prominence in the evaluation of computer vision models, particularly for applications that are designed to run in the real-world.

This detailed analysis is designed to give a clear overview, making it easier to understand how these metrics work both on their own and together. This will help in evaluating the overall performance of models designed for image segmentation.

3.2.1 Intersection of Union (IoU) and Mean Intersection of Union(mIoU)

The Intersection of Union (IoU), or Jaccard index, developed by Grove Karl Gilbert, serves as a standard to evaluate the likeness of the predicted mask segmentation with the ground truth mask [30]. The computation involves dividing the number of common pixels (True Positives) between the two masks by the total number of pixels. If the IoU

Table 1 This table provides a comprehensive overview of publicly available datasets for image segmentation tasks, categorizing them based on essential characteristics. The segmentation tasks suitable for each dataset are outlined, specifying whether the dataset is applicable for Instance, Semantic, and/or Panoptic Segmentation. The type of images within each dataset, the available annotations, the total number of images, and the total number of classes are presented. In the task column, **Image Segmentation** indicates datasets suitable for various segmentation tasks. * indicates information not reported. — indicates nonexistence

Datasets	Tasks	Scenes	Type of Annotations	Total number of images	Total number of classes
Cityscapes [11]	Image Segmentation	Street	Polygonal Bounding Box	25 000	30
COCO-stuff [3]	Image Segmentation	General	Pixel-level semantic	164 000	172
VOC [13]	Image Segmentation	General	Instance Semantic ROI	23 060	20
MVD [17]	Image Segmentation	General	Polygonal Semantic	25 000	124
GTA 5 Synthetic Images [18]	Semantic Segmentation	Synthetic Street	Mask	25 000	19
LVIS [19]	Instance Segmentation	General	Mask	164 000	1200
KITTI [14]	Image Segmentation	Street	Bounding Box	*	30
KINS [20]	Instance Segmentation	Street	Instance Semantic	15 000	13
Open Image [21]	Instance Segmentation	General	Bounding Box Instance	9 000 000	600
CamVid [15]	Semantic Segmentation	Street	Semantic	700	32
ADE20K [23, 24]	Semantic Segmentation	General	Polygonal	27 000	150
SA-1B [25]	Image Segmentation	General	Mask	11 000 000	—
OPPD [22]	Instance Segmentation	Plant	Bounding box	7 590	47
IDD [26]	Semantic Segmentation	Street	Pixel-level semantic	10 004	34
BDD100K [27]	Semantic Segmentation	Street	Polygonal Mask	100 000	40
ACDC [28]	Instance Segmentation	Street	Pixel-level semantic	4006	19
IDD-AW [29]	Semantic Segmentation	Street	Pixel-level semantic	5 000	30
	Instance Segmentation				

is higher than a predetermined threshold, it is considered a True Positive; otherwise it is a False Positive. To help clarify this concept, Figure 2 provides an example.

Equation 1 presents this metric using the previous concepts,

$$IoU = \frac{TP}{TP + FP + FN}, \quad (1)$$

where TP stands for true positives, which is calculated as an AND, as seen in Equation 2,

$$TP = GT \times P, \quad (2)$$

Ground truth (GT) and prediction mask (P) are used in Equation 3 to identify false positives (FP).

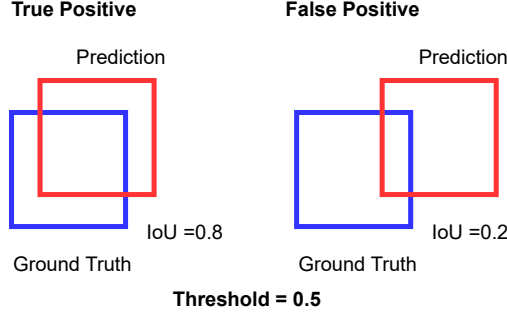


Fig. 2 Example of True Positive and False Positive with a threshold of 0.5

$$FP = (GT + P) - GT, \quad (3)$$

Equation 4 shows how false negatives (FN) are calculated, which are all the pixels the model failed to predict.

$$FN = (GT + P) - P. \quad (4)$$

Equation 5 represents the intersection and union concepts applied to each image class.

$$IoU = \frac{GroundTruth \cap Prediction}{GroundTruth \cup Prediction} \quad (5)$$

Afterward, the final result or mIoU result is determined by taking an average of the results obtained. It should be noted that the mIoU metric is highly sensitive to imbalanced classes. To put it simply, if some classes are more prevalent than others, they have a greater influence on the overall mIoU score. On the other hand, if there are only a few classes and they are misclassified, it can significantly impact the mIoU score. In other words, errors in the classification of rare classes have a greater impact on the mIoU score than the misclassification of more common classes. This can be especially problematic in imbalanced datasets.

3.2.2 Pixel Accuracy

A simple metric used mostly in semantic segmentation is Pixel Accuracy [31]. This metric measures the number of pixels that were correctly classified. Furthermore, this evaluation can be made class-wise and across all classes. Equation 6 represents per-class pixel accuracy, which normally uses a binary mask.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (6)$$

where TP stands for True Positives, TN for True Negatives (correctly classified pixels that do not represent a class), FP for false positives, and FN for false negatives.

Additionally, the overall accuracy can be calculated by finding the ratio of correctly identified pixels to all classes, regardless of the class. The accuracy is reported in percentage. Although pixel accuracy is a straightforward metric, it has some limitations. One of its main drawbacks is that misalignment between the ground truth and the prediction can significantly affect the accuracy of the metric. Moreover, the method disregards the spatial structure of an object, thereby failing to assess the model’s ability to accurately capture the object’s shape and boundaries.

3.2.3 Average Precision (AP)

The Average Precision [32] metric is a commonly employed evaluation measure in the task of instance segmentation. Given that the instance segmentation task involves the generation of multiple masks for each class, assessing the performance of a model becomes more challenging as compared to semantic segmentation. As a result, AP metric combines two concepts of machine learning: precision, which measures the accuracy of positive predictions, and recall, which assesses the percentage of positive predictions. Equations 7, 8 show how both concepts are calculated,

$$Precision = \frac{TP}{TP + FP}, \quad (7)$$

$$Recall = \frac{TP}{TP + FN}, \quad (8)$$

where TP means true positives, FP is for false positives, and FN is for false negatives.

A precision-recall curve is utilized to evaluate. To generate a precision-recall curve, it is necessary to vary the threshold. The Average Precision (AP) is calculated as the area under this curve, providing a comprehensive measure of the model’s performance in balancing precision and recall. The curve is then averaged across different precisions at several recalls or an interpolation of precision values.

3.2.4 Panoptic Quality (PQ)

The Panoptic Quality metric, as described in the research paper by [33], assesses the effectiveness of panoptic masks. This metric is designed to address the limitations of evaluating things and stuff separately by combining both segmentation types into a single score. The calculation process for this metric is outlined in equation 9.

$$PQ = \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|} \times \frac{\sum_{(p,g) \in TP} IoU(p,g)}{|TP|} \quad (9)$$

The product comprises two distinct components: detection quality (the first fraction) and segmentation quality (the second fraction). Detection quality measures the

extent to which the model accurately matches object instances between the ground truth and the segmentation mask. It's important to note that the evaluation of false positives and false negatives, with a weightage of 1/2, determines the balance between accurate detection and precise segmentation. The second component, segmentation quality, evaluates how well the segments align with the ground truth by using the Intersection over Union metric. The final output, known as Panoptic Quality, ranges from 0 to 1, with 1 indicating a flawless panoptic segmentation.

3.2.5 Speed (FPS)

The ability of an image segmentation method to process quickly is crucial for determining its practical use. To evaluate this, the FPS (frames per second) metric is employed. This metric is calculated by dividing the total number of processed frames by the time taken to process them. A lower value indicates a slower processing rate, while a higher value indicates a faster processing rate. This metric is becoming increasingly used in real-time segmentation model evaluations. This metric has become an important tool in evaluating the effectiveness of real-time segmentation models, which must process a minimum of 30 FPS frames per second to qualify as such [34].

4 Key Concepts in Image Segmentation Architectures

Image segmentation and computer vision are grounded in fundamental concepts that serve as the building blocks of models and architectures. To understand how authors create innovative state-of-the-art models that push the boundaries of precision, it is crucial to comprehend these pillar concepts. Gaining this understanding is a prerequisite for grasping how authors strategically employ these concepts to construct their models. Therefore, this section provides a brief explanation of these concepts.

4.1 Fully Connected Layer or Dense Layer

The fully connected layer or dense layer is a building block in artificial neural networks. This type of layer connects each neuron in the current layer to every neuron in the previous layer, creating dense connections between them. These connections are represented by weights, usually updated using the backpropagation algorithm. Backpropagation computes gradients by propagating the cost function's information backward through the network. This process enables the network to learn and improve its performance.

The representation of a dense layer is shown in Equation 10,

$$y = f(b + W * x), \quad (10)$$

where W is the weight matrix, where each element W_{ij} represents the weight of the connection between the neuron i in the current layer and the neuron j of the previous layer. The x represents the input vector, b stands for the bias vector, and f is the activation function, which is applied element-wise to the result of $b + Wx$.

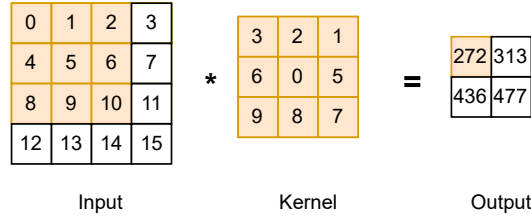


Fig. 3 Illustration of a convolution operation, where the input is a tensor, which is matrix 4x4, and the kernel is matrix 3x3. The filter will go through the matrix, creating a matrix 2x2 where the entries are the results of each convolution operation

4.2 Convolutional Neural Networks (CNN)

Convolutional neural networks (CNN) were the algorithm that enabled a breakthrough in the world of computer vision. The term convolutional networks comes from the fact that these networks use a mathematical operation called convolution in at least one of their layers, rather than a general matrix multiplication like other neural networks. Convolution is a linear operation that uses two functions of real value to create a third function that shows how the shape of one function is changed by the other, as shown in Equation 11:

$$s(t) = (x * w)(t), \quad (11)$$

where $*$ denotes the convolution operation. The first argument (the function x) to the convolution is called the input, and the second argument (the function w) is called the kernel. The output is known as the feature map.

It is important to note that a normal input of CNN are tensors and kernels are composed of a multidimensional array of parameters that are adapted by the learning algorithm. Besides that, usually convolution operations are practice in more than one axis at the time, thus the equation is now represented as follows:

$$s(i, j) = (K * I)(i, j) = \sum_m \sum_n I(i + m, j + n) K(m, n). \quad (12)$$

Equation 12 is a representation of a function called cross-correlation, which works the same way as convolution. Neural networks, in practice, use this method instead of the actual convolution that requires flipping the kernel. Figure 3 illustrates an explanation of how basic convolution works in neural networks.

Furthermore, convolution holds three important concepts that improve neural network learning: sparse interaction, parameter sharing and equivariant representations.

In CNN, the interactions between input and output are far less due to the size of the kernel that is usually set to be smaller than the input, hence the name "sparse interaction". Moreover, fewer connections mean a lower number of operations needed

and thus less computation required. Additionally, fewer interactions also mean less parameters to be stored, reducing the memory needed. Although it is important to notice that in CNN this method does not affect runtime but decreases significantly the memory requirements of the model, by having far less parameters to store. A particularity caused by parameter sharing on a convolutional layer is a property called equivariance. Equivariance means if the input changes, the output also changes. Translated to convolution neural networks terminology, this means that if the input is moved some amount to the right, then its representation will move the same amount to the right. This is advantageous when trying to detect edges of an object in an image. Due to the edges appearing more or less on the same places on an image, using shared parameters across the entire image becomes really useful.

There are three stages: convolution stage, detector stage and pooling stage. First of all, a set of linear activations is produced by the layer which operates multiple convolutions in parallel. Secondly, each linear activation is the input of a nonlinear activation function, for example, the ReLU activation function. Lastly, with the help of a pooling function, the output of the layer is modified even further. The pooling function replaces the output of the network with a summary of the nearby outputs statistics. This means that by using, for instance, a max pooling operation, the output will be the maximum output within a rectangular neighborhood. There are several other popular pooling functions, but all of them help the invariance of the model to small translations of the input. Where invariance to translations means that the pooled output is unchanged by the translation of the input by a small amount. Moreover, if the goal is to know if a feature is present, invariance to local can be really helpful.

Another interesting aspect is the fact that by pooling over the outputs of convolutions that were separately parametrized, the features can learn which transformations to be invariant to. Additionally, it is possible to use fewer pooling units than detector units due to the fact that pooling summarizes the response over a whole neighborhood. Hence, improving computational efficiency of the network, making it faster. On the other hand, a reduction in the input size when the parameters number in the next layer is a function of its input size, and because of this, the memory requirements are less and the statistical efficiency may be improved.

Pooling is a great tool when handling inputs that have distinct sizes. This is because pooling requires the input to have a fixed size, which is achieved by adjusting the offset between pooling regions. This way the classification layer always receives the same number of summary statistics.

4.2.1 Spatial separable convolution

When defining an image or kernel, the spatial dimensions to consider are the width, height, and depth. Spatial separable convolution exclusively considers the first two dimensions and functions by decomposing a kernel into two smaller ones, which reduces the number of multiplication operations and improves efficiency. Nevertheless, not all kernels can be broken down into smaller ones, which is a crucial factor to consider when deciding which method to use.

4.2.2 Depth-wise separable convolution

Depth-wise separable convolution is a specific type of convolution, brought by [35], that involves breaking down standard convolution into two stages. First, the depth-wise convolution applies a single convolutional filter to each channel, individually conducting spatial filtering for each input channel. After that, a point-wise convolution generates a linear combination of the outputs. This two-stage process reduces the number of parameters and increases efficiency, as the depth-wise convolution is less computationally expensive. Additionally, point-wise convolution helps to combine information across channels and capture cross-channel dependencies. As a result, depth-wise convolution is often used in memory-restrained models, such as those utilized in mobile and edge applications.

4.2.3 Deformable convolution

Computer vision faces a considerable challenge when presented with images containing variations in scale, object deformation, differing points of view, or intra-class variability. To address this problem, [36] introduced deformable convolutions. This technique modifies the convolution operation by adjusting the pixels that the model examines through the addition of an offset, denoted as Δp_n in Equation 13.

$$s(p_0) = \sum_{p_n \in \mathfrak{R}} (w(p_n) * x(p_0 + p_n + \Delta p_n)), \quad (13)$$

where, $s(p_0)$ is the output feature map of the pixel of interest, \mathfrak{R} is the receptive field size and dilation, w is the kernel that represents and lastly the x that is the input where the offset is defined. One important problem that arises is that it is not known whether the defined offset will be inside the defined grid or not. Therefore, bilinear interpolation is necessary to ensure that all the offsets are inside the grid. The concept outlined above can be referenced in the article authored by [36].

4.3 Residual Block Layer

A residual block, introduced by [37], is a short connection unit that allows skipping connections, which in this case is also called identity connections. A representation of this block can be seen in Figure 4.

The most important part of this block is identity mapping, whose sole purpose is to add the output of the previous layers to the following layer. The identity mapping does not contain any parameters, so it is a pure addition operation. Nevertheless, it is essential to be aware of the dimensions of $F(x)$ and x , because in most cases they will not be equal, since convolution operations usually reduce the spatial resolution of an image. Therefore, it is multiplied x by a linear projection W to equalize the dimensions as presented in Equation 14:

$$y = \mathcal{F}(x, W_i) + W_s x, \quad (14)$$

where $\mathcal{F}(x, W_i)$ represents the residual mapping to be learned.

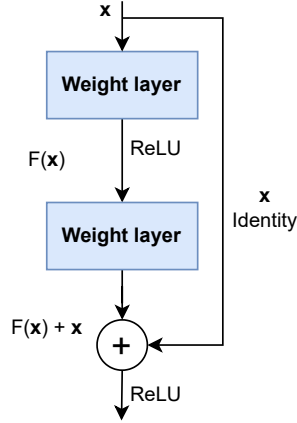


Fig. 4 Representation of a Residual Block

4.4 Attention mechanism

Computer vision architecture designers have taken a keen interest in a new building block in the field of neural networks - attention. This technique is designed to mimic the way humans select important elements in an image. Although attention was created for language translation, it has proved helpful in computer vision by allowing for a better understanding of images. Previously, natural language processing models faced the problem of being unable to memorize long sentences, which led to the output of nonsensical text. To solve this, [38] developed the attention mechanism, which added attention weights to the encoder-decoder process of NLPs. These attention weights were the amount of attention to be paid to a specific hidden encoder unit. The amount of attention weights was equivalent to the hidden encoder units. The hidden weights were combined linearly using the attention values to create the novel context vector. Then, the context vector was computed along with the decoder's previous hidden state and output to generate a new hidden state and output.

[39] introduced the attention mechanism in computer vision tasks for image captioning. Instead of using a fully connected layer, they utilized a VGG [40] and extracted features from a lower convolution layer and flattened it, resulting in a set of annotation vectors, which are a D-dimensional representation of a specific part of the image. These vectors are concatenated to form matrix a , the output of the CNN, which the attention module uses to determine relevant image sections for generating the next word. Hard attention and soft attention are used for this purpose, both requiring the annotation vectors and attention weights. Hard attention separates certain parts of the image that are the ones considered when producing the following word in the caption. On the other hand, soft attention demonstrates the relative importance of each part of the picture to the other parts. The combination of attention weights and

annotation vectors with the function of the hard and soft attention produces the context vector, which is then fed into a long short-term memory (LSTM) along with the previous decoder hidden state and an embedding matrix to output the caption. The attention weights are calculated using the MLP’s energy score, which is then fed into a softmax to obtain the final attention weight.

4.4.1 Global attention and local attention

[41] introduced two distinct attention modules. The global attention module takes into account all the elements in the input sequence to generate each element of the output sequence. It assigns varying weights to each input element, depending on its relevance to the decoding step at that moment. These weights determine the importance of each element in producing the current output. While this module captures long-range dependencies and context, it also involves higher computational costs. In contrast, the local attention module concentrates on a specific, limited section of the input sequence at a time, which is more efficient. For each decoding step, this module selects a subset of the sequence, known as a local window. It then calculates attention weights for the elements within the window, similar to the global attention weights. Each weight denotes the importance of each element in the current decoding step. Based on these weights, a context vector is computed, which influences the generation of the current output. As the decoding process progresses, the local window shifts to another part of the sequence, enabling the model to attend to different sections of the input in distinct decoding steps. This module has been particularly helpful for long input sequences, allowing the model to focus selectively on relevant portions and capture local dependencies efficiently.

4.4.2 Channel attention

[42] presented an innovative channel attention module in their research paper, which aims to pinpoint the channels that have significant information. The module utilizes both averaging pooling and max pooling techniques on a feature map and subsequently processes the output through an MLP. The MLP generates two vectors, one for max pooling and one for average pooling, which are then combined and passed through a sigmoid linear function to derive the channel attention.

4.4.3 Spatial attention

[42] proposed the concept of spatial attention in their research, which aids in identifying where important information lies on the spatial axis. The process entails subjecting a feature map to max and average pooling, followed by the concatenation of the results. The resulting output is then fed through a convolutional layer to detect the most critical areas, followed by a sigmoid function that yields the spatial attention value ranging from 0 to 1.

4.4.4 Self-attention

Self-attention, first introduced by [43], can be broken down into four steps. To begin, a neural network layer, called positional encoding, is used to encode the position

information that captures the relative relations between pixels in the image. Initially, the mechanism extracts a query, which is essentially the information it seeks. It then proceeds to extract the relevant details from an image that may be associated with the Query. Ultimately, it retrieves the precise item it is seeking, which is referred to as the value. Attention leverages the positional encoding and applies a neural network layer to transform it and produce the Query (Q). A different neural network layer is then applied to the same positional encoding to generate the Key (K). This operation is used once again, with a different neural network layer, generating the Value (V). Once these components have been derived, attention compares them to determine where the neural network should focus its attention within the image. This is accomplished by calculating the attention score, which represents the similarity between the Query and the Key using the cosine similarity formula (Equation 15).

$$sim(Q, K) = \cos(\theta) = \frac{Q * K}{\|Q\| \|K\|}, \quad (15)$$

which is used a dot product between two vectors in the numerator and then multiplied their norms in the denominator to normalize the value between -1 and 1. This equation can be adapted for matrices, as shown in Equation 16.

$$Similarity(Q, K) = \frac{Q * K^T}{scaling}, \quad (16)$$

which is often called the Similarity metric. This Similarity metric is a crucial point to get attention weighting, which is the model finding where to attend in the image by computing how similar the Query is to the Key. The attention weighting is calculated as follows (Equation 17):

$$softmax(\frac{Q * K^T}{scaling}), \quad (17)$$

which outputs a relative score value between 0 and 1. Using this metric, the self-attention mechanism extracts important features by multiplying attention weightings with the Value as shown in Equation 18,

$$A(QKV) = softmax(\frac{Q * K^T}{scaling}) * V. \quad (18)$$

The concept of the self-attention mechanism can be better understood with the aid of Figure 5. The representation depicts a single head that produces a set of features that are given high attention. Neural network architectures can have several such heads, called multi-headed attention, that extract different features with high attention. These features represent different relevant parts of an image, which helps in generating a richer representation of the presented data.

4.5 Transformer

The concept of the transformer was first introduced by [43] when working with translation mechanisms. The authors relied on attention mechanisms, specifically

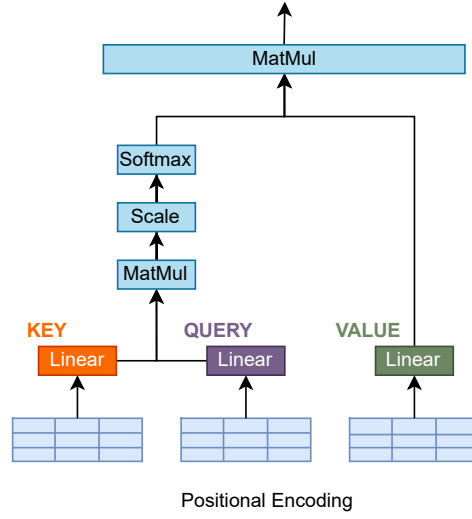


Fig. 5 Illustration of a self-attention head

self-attention and position encoding to create this model. The strategy they used revolutionized the world of deep learning, which has since used the Transformer method for all sorts of things, including computer vision.

Before the encoder section, the input words are transformed into a vector by the input embedding, while the positional encoding is used to address the sequential nature of language. The encoded words then pass through the encoder layer, which has two sub-modules: a self-attention mechanism to capture dependencies between words efficiently and a fully connected layer. The encoder layer has residual connections integrated into its architecture to mitigate the vanishing gradient problem, a common challenge in deep networks. The residual connections also selectively retain relationships among words while discarding redundant information from both word embedding and position encoding. Similarly, the decoder generates vectors representing the target language through the use of output embeddings and positional encoding. Then, a multi-head attention mechanism is employed to allow the model to attend to different positions in the input sequence, enabling it to capture relevant information from the entire source sequence. Additionally, the first multi-head block is utilized for encoder-decoder attention. During this process, the encoder's output provides keys, while the decoder's position generates queries. These are compared to determine the importance of different parts of the input sequence, represented as values. By employing residual connections, the encoder-decoder attention can focus on the relationships between output words and input words without the need to retain word embedding and position encoding that occurred earlier. The output of the multi-head block is

then smoothly passed to a feedforward layer, culminating in a softmax activation that accurately detects and outputs the translated phrase.

4.5.1 Vision Transformer

[44] developed the Vision Transformer, a seamless integration of Transformers in the Computer Vision world.

The use of Transformers in computer vision has encountered a significant obstacle in dealing with input sizes. Due to attention being a quadratic operation, processing even a small image with current hardware has become infeasible. To address this issue, the authors have introduced global attention by breaking down an image into patches and transforming them into a sequence. They then used position embedding to inform the network about the positional sequence and passed the patch through a linear projection. The resultant unrolled patch (vector) was then subjected to multiplication with an embedding matrix, and the product was forwarded to the standard encoder transformer in conjunction with the position embedding. The encoder transformer receives a unique input, which is a learnable embedding, which subsequently undergoes classification through a multi-layer perceptron, ultimately producing an output class.

5 Common Deep Network Architectures

Numerous developments in the area of image segmentation are built on the architecture of well-known backbone networks. This section will explore a selection of deep networks that are widely recognized for their significant contributions to the field.

5.1 AlexNet

AlexNet [45] was the first to dismantle traditional computer vision techniques and introduce CNN to computer vision. Krizhevsky et al. achieved breakthrough results at the ImageNet Large Scale Visual Recognition Challenge, *ILSVRC-2012* with a top-5 test accuracy of 84.6%. Its architecture contains five convolutional layers, which include response-normalization layers after the first and second convolutional layer and max-pooling layers at the end of these layers and at the end of the fifth convolutional layer. In addition, three fully connected layers (FCN) are added, where a dropout is used after the first two layers, to reduce the probability of overfitting. Both the convolutional layers and the connected layers are followed by the activation function ReLu.

5.2 VGG

The VGG architecture was developed by the Visual Geometry Group, which is part of the Department of Science and Engineering at Oxford University [40]. This architecture ranges from 16 convolutional layers (VGG-16) to 19 convolutional layers (VGG-19).

The improved properties of this model lie in the fact that they are deeper than previous models. This is achieved by grouping a stack of convolutional layers that use 3×3 receptive fields. These are the smallest filters capable of capturing left/right and

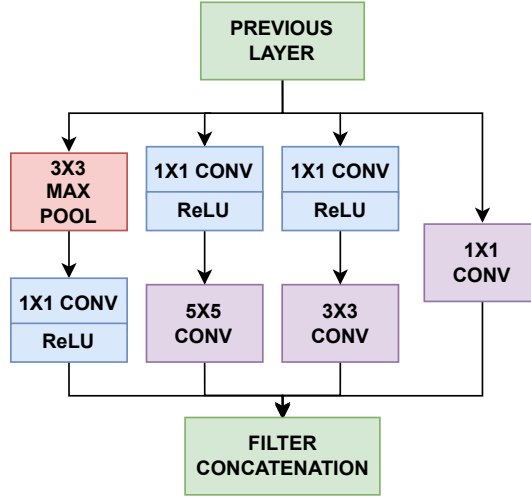


Fig. 6 Representation of the inception module with dimension reductions used on GoogLeNet

up/down movement. These filters are followed by three fully connected layers. A 1x1 convolutional layer is integrated into the stack of convolutional layers in the middle to increase non-linearity without affecting the filters of the previous convolutional layers. Previous work usually worked with larger filters in the convolutional layers and at the same time with fewer layers, e.g. with only one fully connected layer, which required more parameters and increased linearity, making it harder to train and possibly leading to overfitting. The VGG-16 achieved 92.7% top-5 test accuracy on the *ILSVRC-2013*.

5.3 GoogLeNet (Inception-v1)

[46] presented a novel architectural approach, drawing inspiration from the work of [47]. This new design introduced an inception module into the pipeline, as depicted in Figure 6. The inception module revolutionizes the stacking of convolutional layers, allowing for a more diverse configuration. It comprises a network-to-network (NiN) layer, a pooling operation, and two parallel convolution layers, with one being larger than the other. Subsequently, 1x1 convolution operations are applied to compute reductions, followed by the application of rectified linear activation (ReLU) to the results. The GoogLeNet network stands out due to its complexity, incorporating a total of 22 layers. Notably, it achieved impressive results during the *ILSVRC-2014* challenge, boasting a top-5 test accuracy of 93.3%.

5.3.1 Inception-v2

Inception-v2 [48], an updated version of GoogLeNet [46] architecture, was developed by Szegedy et al. One of the main drawbacks of the original GoogLeNet was a decrease

in accuracy when using larger convolutions such as 5×5 or 7×7 , which led to a reduction in dimensions and potential loss of information. To address this issue, Szegedy et al. introduced factorization of these larger convolutions into multiple 3×3 convolutions. This not only increased computational efficiency but also improved overall performance. Furthermore, the $n \times n$ convolutions were decomposed into a combination of $1 \times n$ and $n \times 1$ convolutions, resulting in a 33% reduction in computational cost. In line with the objective of creating a wider pipeline rather than a deeper one, the filter banks in the inception module were expanded to alleviate representational bottlenecks.

5.3.2 Inception-v3

In the case of Inception-v3, [48] observed that the auxiliary classifiers acted as regularizers. To enhance the performance, they incorporated batch normalization into these auxiliary classifiers. Additionally, the authors introduced the use of the RMSProp Optimizer and implemented Label Smoothing. Another modification involved factorizing the 7×7 convolutions. These changes in Inception-v3 yielded improved results compared to previous state-of-the-art architectures. On the ILSVRC-2012 test set, Inception-v3 achieved a top-5 accuracy of 96.42%.

5.3.3 Inception-v4

[49] conducted a review of previous models and found that they were unnecessarily complicated. As a result, the authors decided to uniformize the architecture. They adjusted the initial set of operations before the inception blocks (stem). These adjustments involved parallel operations that were then fed into a filter concatenation, as opposed to the linear stem of previous models. Additionally, they introduced three inception blocks, denominated A, B, and C, and implemented reduction blocks to modify the height and width of the grid.

5.3.4 Inception-ResNetv1 and Inception-ResNetv2

ResNet [37] inspired [49] to create a hybrid inception model that incorporated residual connections. This resulted in two versions, v1 and v2, with v1 being less computationally expensive than v2. Both versions have the same A, B, and C structure and reduction blocks but with different hyperparameters. The residual connections replaced the pooling layer in the A, B, and C blocks, and to accommodate this change, a 1×1 convolution was added before the summation to match the depth size. The authors also observed that if the number of filters exceeded 1000 in deeper residual units, the model performance suffered, so they scaled the residual activations between 0.1 and 0.3. Additionally, batch-normalization was only used in the traditional layers.

5.4 ResNet

When dealing with Deep Convolutional Neural Networks, it is easy to fall into the belief that more layers produce better results. This belief is based on the common knowledge that by adding layers, the model learns more complex features and thus

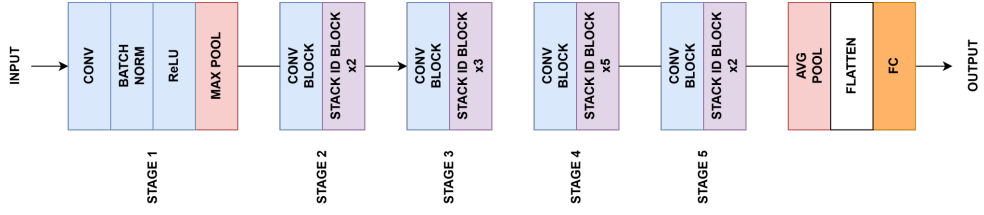


Fig. 7 Representation of a Resnet where four stages of residual layers with multiple identification mapping blocks are found

learns better and produces better results. And if for some reason the results stop improving, the blame should be put on overfitting [37]. In this case, the solution is to add regularization parameters such as L^2 or dropout. However, it has been proven that even Deep Learning models can have a maximum depth limit after which the models no longer produce better results [50]. This consequence can have several reasons, such as initialization of the network, optimization, and the most common reason, the vanishing gradient problem. However, using batch normalization ensures that the gradients have expected norms. Therefore, studies have shown that adding a layer called Residual Block smooths the problem. The use of residual blocks in a neural network was first introduced by [37] and is what constitutes a Residual Neural Network, known as ResNet.

Figure 7 shows an illustration of an implementation of a Resnet50 on a CNN for plant identification problem [51]:

This extension of Convolutional Neural Networks allowed the training of much deeper networks and increased performance on multiple different tasks such as semantic segmentation. To prove this concept, the ResNet-152 was tested on the *ILSVRC-2015* where it achieved a 3.57% top-5 error, e.g. 96.4% top-5 accuracy, far better than previous pipelines.

5.5 ResNeXt

Inspired by the ResNet structure of repeating layers, [52] designed ResNeXt, an architecture constituted by a stack of residual layers in which each layer contains a group that aggregates a set of transformations. This group is called *cardinality group*, where the term cardinality indicates the number of parallel paths within a group. Each path is responsible for learning a different feature. These groups are implemented as bottleneck blocks, consisting of 1x1 convolution, 3x3 convolution, and another 1x1 convolution. The outputs of each path are then aggregated together to create the group's final output. The fact that each layer can have multiple paths allows the network to learn a more diverse set of features. Furthermore, by increasing the depth and width simultaneously, ResNeXt uses more parameters, which permits better performance. To evaluate its performance, ResNeXt was tested on the *ILSVRC-2016* validation set where it achieved 4.4% top-5 error, e.g., 95.6% top-5 accuracy.

5.6 MobileNet

MobileNet [53] was introduced to fulfill the need for an architecture suitable for deployment in embedded devices such as mobile phones and edge devices while still delivering decent accuracy and real-time performance. The framework is characterized by a sequence of depth-wise separable convolution layers, comprising a depth-wise separable convolution layer followed by batch normalization, the application of the ReLU function, and a 1x1 convolution layer. Only the initial layer is a simple convolutional layer, and the conclusion of the pipeline is marked by an averaging pool layer followed by a fully connected layer, resulting in a total of 28 layers in MobileNet. Moreover, the authors introduced two key hyperparameters to improve the adaptability of MobileNet across different devices. The width multiplier uniformly reduces the number of filters in each layer, thereby creating a thinner and computationally more efficient network capable of higher speed. Additionally, the resolution multiplier offers a means to reduce computational cost by adjusting the input resolution. MobileNet achieved 70.6% accuracy on the *ImageNet* classification task.

5.6.1 MobileNetV2

MobileNetV2 [54], was built upon the foundation of MobileNet, aimed to reduce computational costs further while improving performance. It introduced several key changes to achieve this. Firstly, it replaced depth-wise separable convolution layers with bottleneck depth-separable convolution layers. Secondly, it incorporated residual connections in the narrower parts of the network, resulting in what they termed an inverted residual structure. This design enables the compression of information, potentially leading to information loss. To address this, Sandler et al. introduced the concept of linear bottleneck, where the last layer in each block has a linear output to prevent excessive information loss by non-linearities. This facilitates gradient propagation across multiple layers, improving memory efficiency and improving the training process. Additionally, the ReLU6 activation function was introduced, and a uniform expansion of the hidden state in each residual was performed to balance computational cost and model performance. These changes allowed the model to be applied in semantic segmentation. When tested on *ImageNet* classification task, MobileNetV2 achieved a top-1 accuracy of 74.7%.

5.6.2 MobileNetV3

[55] developed MobileNetV3, an updated version of MobileNet. To improve channel attention, the authors implemented a Squeeze-and-Excitation (SE) module, allowing the model to focus on more important features. These modules were integrated into specific layers of the bottleneck block. Additionally, the authors incorporated a Hard Swish activation function to improve the model's non-linear capabilities and adjusted the expansion ratio in specific bottleneck layers. To ensure better adaptability to various edge devices while balancing performance and computational cost, the authors utilized a platform-aware neural architecture search (NAS). These modifications reduced the parameter count and latency, making the model lighter and faster

while demonstrating improved performance. MobileNetV3 was assessed on *ImageNet* classification task, achieving a top-1 accuracy of 75.2%.

5.6.3 MobileNetV4

[56], built a family of models from the MobileNetV4 framework. They implemented various improvements, starting with the Universal Inverted Bottleneck (UIB), an extension of the inverted bottleneck (IB). This module introduced two optional depth-wise convolutions, with one preceding the expansion layer and another following the projection layer, both of which are determined by the NAS optimization procedure, providing the architecture with significant flexibility. Additionally, the authors considered four distinct blocks: the original inverted bottleneck block, the ConvNext block, the FFN block, and a novel Extra depthwise IB (ExtraDW) block that can be instantiated on the UIB. Thus, UIB facilitates dynamic spatial and channel mixing adjustments, receptive field size, and computational resource usage, optimizing network performance for diverse tasks and conditions. Furthermore, inspired by large language models, the authors introduced a multi-query attention (MQA) module, which uses shared keys and values across all heads, improving memory efficiency. The MobileNetV4 was evaluated on *ImageNet-1K* classification task, achieving 83.4% top-1 accuracy.

5.7 Xception

[57] introduced a new pipeline called Xception, which obtained 94.5% on the validation set in the *ILSVRC-2012* challenge. Chollet observed that an extreme case of the Inception module can be compared to a depth-wise separable convolution. While the Inception module and the depth-wise separable convolution share similarities, there are some differences. In the depth-wise separable convolution, the 1×1 convolution is done after the channel-wise spatial convolution. Furthermore, depth-wise separable convolution lacks the ReLU layer. Xception, which stands for Extreme Inception, comprises 36 convolutional layers organized into 14 modules and replaces the Inception module with the depth-wise separable convolution. Furthermore, linear residual connections are employed between the modules, excluding the first and last modules. These connections help gradient propagation and information flow throughout the network.

5.8 Swin Transformer

The Swin Transformer [58] is the first suitable general backbone that introduces transformers to computer vision tasks. This model is capable of hierarchical feature maps, which permits the use of techniques such as feature pyramid networks or other models specialized in segmentation. Furthermore, it uses the window shift mechanism, which causes the window partition to move between successive self-attention layers building connections between them. Because the number of patches in each window is fixed, the complexity becomes linear to the image size.

5.9 Visual Attention Network

A novel linear attention mechanism that uses kernels is presented by [59]. This module is embedded into the self-designed Visual Attention Network (VAN) neural network. The large kernel attention module is able to adapt to both channel and spatial dimensions by decomposing larger kernel convolution steps. This enables self-attention and large kernel convolution, important aspects for computer vision tasks. The VAN network is a four-stage hierarchical structure, where each stage performs batch normalization twice, and uses an attention mechanism and a feedforward network (FFN) to extract the features.

5.10 Vision Transformer with Deformable Attention

[60] were inspired by the Deformable Convolution Networks [36] and developed a similar mechanism for Vision Transformers. They introduced a new backbone called Deformable Attention Transformer (DAT), that employs a deformable multi-head attention module (DMHA), to enable the model to learn deformed key locations for all queries. This approach provides DAT with greater flexibility and efficiency in capturing informative features. The model was evaluated as the backbone for Mask-RCNN [61] and Cascade Mask-RCNN [62] using the *COCO* dataset, achieving scores of 44.0% AP and 45.8% AP, respectively. [63] later created an improved version of DAT, called DAT++, which simplified the design and removed several hyperparameters. Furthermore, they extended the DMHA module to all stages of the model, removed the constraint range factor on the learned offsets, and unified the number of deformed keys across stages. Additionally, they implemented an overlapped patch embedding and downsampling technique to enhance the model’s ability to capture local features and positional information. DAT++ was then tested as the backbone for the same models, achieving an average precision of 45.7% for Mask-RCNN [61] and 47.0% AP for Cascade Mask-RCNN [62].

5.11 ConvNeXt

Vision Transformers have been the breakthrough of computer vision in the last few years, outperforming convolutional backbones. This reality motivated [64] to design a new convolutional backbone with features inspired by vision transformers. To design ConvNeXt, Liu et al. used ResNet [37] as the base and trained the architecture with a similar training method as the Swin Transformer [58], such as the AdamW optimizer [65], regularization schemes and augmented data methods. In the second stage, several changes were made to the base ResNet architecture. First, they modified the number of blocks in each stage from (3,4,6,3) to (3,3,9,3) and used a patchify stem, a 4x4 non-overlapping convolution, at the beginning of the network. They also implemented depth-wise convolution in combination with 1x1 convolution layers to achieve separation of spatial and channel mixing, which improves the network’s performance. Additionally, the inverted bottleneck design was adopted, and a 7x7 kernel size was used for the convolution in each block. Liu et al. replaced the commonly used RELU activation function with a Gaussian Error Linear Unit (GELU), used only once in each block. Moreover, instead of batch normalization, Layer Normalization (LN) was

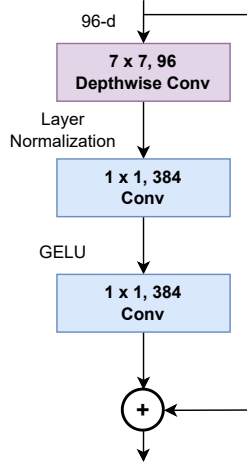


Fig. 8 Representation of a ConvNeXt block design as proposed by [64]

implemented and executed only before the 1×1 convolutional layers. Figure 8 shows the ConvNeXt block. In a final remark, Liu et al. used separate downsampling layers constituted of 2×2 convolutional layers with the stride of 2 and multiple LN layers. These modifications lead ConvNeXt to achieve 87.8% on ImageNet top-1 accuracy.

6 Semantic Segmentation

In the following section, the main advancements in semantic segmentation until 2024 are described. The organization closely follows the structure of [66], in which the methods are organized depending on their underlying base features. Table 2 reports the scores of the methods discussed on this section.

6.1 Fully Convolutional Networks

The materialization of fully convolutional networks marked a key turning point in image segmentation, proving that deep networks can indeed be trained to effectively segment 2D images of varying dimensions. Fully Convolutional Networks are computationally efficient, have end-to-end learning, and can be used in transfer learning. However, they struggle with complex contextual relationships and produce imprecise boundary masks, which can affect their performance.

[67] were the pioneers in introducing the concept of the fully convolutional network in 2015. By exclusively utilizing convolutional layers, an FCN demonstrates the ability to accept inputs of diverse sizes and generate corresponding segmentation

maps (heatmaps) of equivalent dimensions. The initial step involved replacing fully connected layers in prominent classification networks such as AlexNet [45] and VGG [40]. This adaptation allowed these networks to take in inputs of varying sizes while generating spatial segmentation maps instead of conventional classification scores. Moreover, the integration of skip connections for the purpose of up-sampling feature maps from final layers was of significant importance. This approach ensures that the output retains the same dimensions as the input and imparts the network with pixel-level awareness. These up-sampled maps are fused with feature maps from the initial layers, thereby enabling the network to obtain both features and semantic information. This combined information enables the network to produce accurate semantic segmentation outputs. During evaluations on the *PASCAL VOC 2011* and *PASCAL VOC 2012* datasets, the FCN yielded mIoU values of 62.7% for the former and 62.2% for the latter.

In order to improve the FCN [67] outputs, [68] introduced global context which was not used by the FCN before creating a less accurate solution. ParseNet operates by generating a context vector by pooling the feature map of a specific layer across the entire image. Following this, the context vector undergoes L2 normalization. Subsequently, the normalized context vector is unpooled and fused with the standard feature map, both of which possess identical dimensions. On evaluations with the *PASCAL-Context* dataset, ParseNet scored a mIoU of 40.4%, while on the *PASCAL VOC 2012* dataset, it achieved a mIoU of 69.8%.

6.2 Encoder-Decoder Based Methods

Encoder-decoder models are two-stage architectures that first analyze the prominent features of an image and then leverage that information to generate feature maps, which lead to semantic segmentation masks. The encoder uses various filters and layers to extract the essential features and reduces them to create a smaller representation of the image. The decoder then refines this information, making it larger and generates a detailed map that indicates the location of every object in the image. Using encoder-decoder techniques can be highly effective for comprehending context and capturing objects at various scales while remaining adaptable and straightforward. However, it is important to handle them with care due to the difficulty of tuning the hyperparameters, the necessity for a significant amount of data, and the potential for high computational costs depending on the resolution of the images they are processing.

[69] developed a new pipeline with a primary focus on handling biological microscopy images. The main difference of this architecture is that it consists of two main components, the contracting part, which uses a structure similar to an FCN and employs a 3x3 convolution responsible for feature extraction, called down-sampling, and the expanding part or up-sampling, which uses deconvolution to increase the width and height of the feature maps while decreasing their number. Furthermore, there is an incorporation of concatenated feature maps from the down-sampling stage, which are replicated within the up-sampling segment. This strategic approach safeguards against the loss of pattern information. The culmination of this process is the generation of a segmentation map through a 1x1 convolutional layer, operating on the

final feature map. Worth noting is the U-Net’s deliberate avoidance of fully connected layers, a choice that effectively trims down the parameter count.

Regarding the use of light decoders, [70] suggested a decoder comprised of two new modules, the Dilated Asymmetric Pyramid Fusion (DAPF) and the Multi-resolution Dilated Asymmetric (MDA). The DAPF uses standard atrous convolutions to create a fixed number of feature maps that then go through layers of asymmetric convolution that, due to its architecture, reduces by half the volume of operations, which decreases the computational cost while improving the learning capacity of the module. On the other hand, the MDA module reduces the number of feature maps by half by concatenating and fusing multi-resolution feature maps. Afterward, the asymmetric convolutional branch uses dilated convolutions to extract the contextual information that is then concatenated and passed through the 1×1 convolution layer to match the number of feature maps. Thus the network becomes richer in contextual information that can be used to get large dilation rates by the atrous convolution layers. On the *Cityscapes* validation dataset, FASSD-Net yielded a 78.8% mIoU.

[71], proposed a lightweight model called PP-LiteSeg, which consists of three modules: encoder, aggregation, and decoder. The decoder block is where the innovation lies, as they have integrated three new modules. The Flexible and Lightweight Decoder (FLD) increases feature spatial size and gradually reduces channels, which can be adjusted depending on the decoder. To efficiently strengthen feature representation, they have proposed a Unified Attention Fusion Module (UAFM), which uses channel and spatial attention. Lastly, the Simple Pyramid Pooling Module (SPPM) is integrated to aggregate global context and increase segmentation accuracy. PP-LiteSeg was tested on *Cityscapes* and *CamVid* test sets. PP-LiteSeg has been tested on the *Cityscapes* and *CamVid* test sets. PP-LiteSeg-B, which uses an STDC1 decoder, achieved an accuracy of 77.5% on the *Cityscapes* test set. As for the *CamVid* test set, PP-LiteSeg-B had a performance of 75.0%.

The MetaFormer concept was introduced by [72] as an abstraction of a Transformer. The distinctive architectural choice is the use of an unspecified token mixer. In other words, the design of MetaFormer does not limit itself to attention modules. The authors showed the efficiency of this crucial component by designing a model called PoolFormer, which substitutes the attention module with the pooling operator in a Transformer architecture. This method was used as a backbone by the Mask R-CNN [73] and achieved 37.7% AP on the *COCO* validation set for instance segmentation. On the *ADE20K* validation set, the PoolFormer, equipped with Semantic FPN [74], reached a mIoU of 42.7%. Following the MetaFormer architecture proposed by Yu et al., [75] developed MetaSeg. This framework employs a hierarchical CNN-based encoder to consolidate local information, which is then channeled into a Global Meta Block (GMB) based on the MetaFormer architecture. The GMB integrates the Channel Reduction Attention (CRA) as a token mixer. This multi-head self-attention mechanism captures global information by condensing the channel dimension of the query and key into a single dimension. The features are subsequently processed using a channel MLP merge. When utilizing MSCAN-B [76] as the backbone, MetaSeg achieved mIoU scores of 48.5% on *ADE20K*, 82.7% on *Cityscapes*, and 45.8% on *COCO*.

6.3 Multi-Scale and Pyramid Network-Based Methods

Techniques such as Multi-Scale and Pyramid Networks have been developed to extract information from diverse spatial resolutions and scales in an image. These methods consider context and details across multiple levels of granularity, allowing them to simultaneously focus on both the overall big picture, extracting coarse features and the fine features in more detailed areas. These methods provide better context understanding than others, resulting in excellent performance for semantic segmentation tasks. However, they are complex and require significant computational and memory costs.

In segmentation tasks, real-time applications are highly desired. Yet, some modern semantic segmentation methods prioritize faster inference speed over spatial resolution. To tackle this issue, researchers have conducted multiple studies to improve the efficiency of models without compromising speed or accuracy. [77] developed a model called Image Cascade Network (ICNet) based on a cascade framework. The ICNet consists of three branches, each designed for a different image resolution input. The high-resolution image is downsampled by a factor of two and four. The primary branch is the low-resolution one, which utilizes the PSPNet for semantic extraction. The high and medium-resolution branches then refine the coarse prediction using lightweight CNNs to enhance segmentation quality while significantly reducing the number of parameters. The feature maps produced by each branch are combined using the novel Cascade Feature Fusion (CFF), which involves two feature maps and a ground-truth label to produce the final prediction. Additionally, cascade label guidance is employed during the CFF process to improve the learning process. This ensures that the low-resolution branches guide the higher-resolution branches, allowing the final prediction to capture both global context from the low-resolution branch and detailed local information from the higher-resolution ones. The ICNet was evaluated on the *Cityscapes* and *CamVid* datasets, achieving 70.6% mIoU and 30.3 FPS on *Cityscapes* and 67.1% mIoU and 27.8 FPS on *CamVid*.

The accuracy of Semantic Segmentation can be improved by obtaining high-resolution feature maps through either atrous convolution or feature pyramid fusion. However, the methods used for this purpose can be ineffective or computationally intensive. To tackle this issue, [78] developed the Flow Alignment Module (FAM). This module addresses the problem of misalignment between high-level feature maps fusion and low-level feature maps. The Flow Alignment Module provides explicit and dynamic position correspondence, which is necessary for this problem. Li et al. defined a flow field called Semantic Flow, which is based on the alignment of two adjacent video frames' features in video processing tasks. The Semantic Flow creates a flow of adjacent level feature maps, enabling more flexible fusion and refining low-level features in the semantic representation. SFNet was tested on the *Cityscapes* and *CamVid* datasets. Using ResNet-18 as the backbone, SFNet achieved 80.4% mIoU with 26 FPS on the *Cityscapes* test set. In addition, an evaluation was conducted while performing multi-scale and horizontal flip inference, which achieved 81.8% mIoU on the same test set. Furthermore, for *CamVid*, using DF2 [79] as the backbone improved the baseline by 3.2% mIoU, reaching 70.4% mIoU, and using ResNet-18 achieved an accuracy of 73.8% mIoU.

[80] improved the existing SFNet model by introducing a new flow alignment module called the Gated Dual Flow Alignment Module (GD-FAM). This module combines both high and low-resolution feature maps to generate two semantic flow fields that will warp both feature maps. A gate map is also generated and shared for both flows to complete the fusion process. The researchers tested this new module on two datasets, namely *Cityscapes* and *Mapillary*. Using the ResNet-18 architecture as a backbone, the SFNet-Lite achieved an accuracy of 80.1% mIoU on *Cityscapes*, while the SFNet method achieved 79.8% mIoU at 33.3 FPS. When the STDCv1 net architecture was used as a backbone, it achieved a 78.7% mIoU. On the other hand, when the model was tested on the *Mapillary* dataset using the ResNet-18 and STDCv2 as backbones, SFNet-Lite reached 46.3% mIoU and 45.8%, respectively. Furthermore, the researchers tested the model on a Panoptic Segmentation task by replacing the K-Net [81] backbone and neck for feature extraction while using its head for prediction. This resulted in a PQ of 59.2% with STDCv1 and 60.3% PQ with STDCv2 on the *Cityscapes* dataset.

[82] proposed the Bilateral Segmentation Network (BiSeNet). BiSeNet consists of two paths: the Spatial Path and the Context Path, as well as two modules that enhance accuracy. The Spatial Path uses three convolution layers, while the Context Path utilizes the Xception [57] model as the backbone and an Attention Refinement Module (ARM) to enhance features at each stage. This module easily integrates global context information to produce high-level features. In contrast, the output of the Spatial Path provides low-level features. To combine both outputs, a Feature Fusion Module (FFM) concatenates and balances them using batch normalization. Finally, the results are pooled and computed. Testing using the *Cityscapes* dataset showed that BiSeNet achieved 68.4% mIoU and 105.8 FPS with an Xception-39 backbone. However, when the ResNet-18 was used as the backbone model, it scored 74.7% mIoU.

[83] further improved the BiSeNet network structure for semantic segmentation purposes. They simplified the architecture by removing cross-layer connections and divided it into two branches: the Detail Path and the Semantic Path. The Detail Path has more shallow layers with increased channel usage, while the Semantic Path is redesigned with lighter components. They introduced a new layer, the Guided Aggregation layer, to merge the outputs of both branches, utilizing contextual information from the Semantic Branch to influence the Detail Branch’s feature response. This resulted in the development of a new version of BiSeNet called BiSeNetV2. The performance of BiSeNetV2 was evaluated on the *Cityscapes* dataset, achieving a mIoU of 75.3% on the test set with a frame rate of 47.3 FPS. On the *CamVid* dataset, the larger model achieved a mIoU of 78.5% with a frame rate of 32.7 FPS, while the smaller model, using the Xception-39 architecture as a backbone, attained a mIoU of 65.6% with a frame rate of 175 FPS.

According to [84], it was observed that using a backbone not specifically designed for image segmentation, such as the Xception [57] model, could pose a drawback to the BiSeNet pipeline proposed by [82]. To address this limitation, the researchers designed a new pipeline, depicted in Figure 9, by combining a U-net architecture with a novel module called the Short-Term Dense Concatenate (STDC) module. This module efficiently utilizes fewer parameters while providing different receptive field sizes.

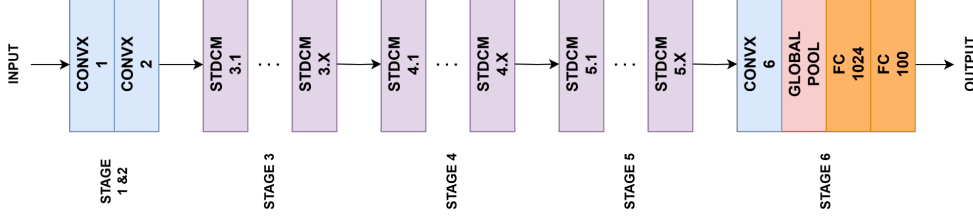


Fig. 9 Representation of the STDC network architecture

Furthermore, the researchers introduced a Detail Guidance module in the decoder phase, incorporating a Detail Aggregation module aimed at preserving spatial details in low-level layers and generating a detailed ground truth. By adopting this module instead of an extra path, they achieved lower computational costs. Testing was conducted on the *Cityscapes* and *CamVid* datasets. In the *Cityscapes* test set, the STDC2-Seg75 framework achieved a mIoU of 76.8% with a frame rate of 97.0 FPS. The STDC2-Seg50 model demonstrated the best speed and performance trade-off, reaching 188.6 FPS with a mIoU of 73.4% on the same dataset. On the *CamVid* dataset, the STDC1-Seg configuration yielded the best FPS value of 197.6 while maintaining a mIoU of 73.0%. The STDC2-Seg model achieved a mIoU of 73.9% with a frame rate of 152.2 FPS on the same dataset.

In order to enhance the inference speed of BiSeNet, [85] proposed some significant modifications. They replaced the Spatial Path with the STDC network as the backbone and introduced two new modules for the decoding phase: the Coordinate Feature Refinement Module (CFRM) and the Coordinate Feature Fusion Module (CFFM), inspired by the Coordinate Attention channel [86]. The Coordinate Feature Refinement Module replaces the ARM module and involves splitting the feature map into two parts: $C \times H \times 1$ and $C \times 1 \times W$. These parts are then encoded along the x and y axes for each channel. A sigmoid function is used to guide the attention vector for the model’s feature learning. The use of multiple channels in the attention vector sets this strategy apart from the traditional approach. As for the Coordinate Feature Fusion Module, it utilizes average pooling to obtain two feature maps, $C \times H \times 1$ and $C \times 1 \times W$, and then feeds each pixel back through gradient backpropagation. By employing permutation and concatenation, a feature map with dimensions $C \times 1 \times (H + W)$ is obtained. After passing through a convolution for channel reduction and a ReLU unit for non-linearity, the feature map is divided into two tensor sums. These tensors are then fed into a sigmoid function, generating an attention vector. This attention vector is multiplied with the original feature map, effectively fusing the features. To further assist the model training, an edge-detecting algorithm, the Sobel operator, is used to output edge data, which improves the features fed to the CFRM, resulting in richer predictions. BiSeNetV3 was tested on the *Cityscapes* test set using four different versions. The BiSeNetV3-75 with STDC2 as the backbone achieved the highest accuracy among all versions, with a mIoU of 79.0% and a frame rate of 93.8 FPS. On the other hand, BiSeNetV3-50 with STDC1 as the backbone core achieved the best frame rate

of 244.3 FPS, with a mIoU of 73.5%. For the *CamVid* dataset, BiSeNet with STDC2 as the backbone reached a mIoU of 76.6% and a frame rate of 147.6 FPS.

Nowadays, in semantic segmentation, it is common to use a two-branch network architecture because of the improvement in its efficiency, especially on real-time tasks. However, this type of architecture suffers from an event in which the small-scale object is overwhelmed by its adjacent bigger objects. [87] names this event as an overshoot and compares it to the overshoot on a PID controller. Regarding this comparison, it was proposed a three-branch network architecture, named Proportional-Integral-Derivative Network or PIDNet, to mitigate this issue. In order to create an architecture that mimics a PID controller in the spatial domain, they add a third branch called the Auxiliary Derivative Branch (ADB). Therefore, the model works with a Proportional Branch which parses and saves the detailed information in its high-resolution feature maps, an Integral (I) branch parses long-range relationships by combining local and global context information and forecasts the border areas, the derivative (D) branch extracts the high-frequency information. An important note is that this architecture requires a precise annotation around the boundary to perform at its best. The PIDNet family was tested on *Cityscapes* and *CamVid* datasets. On the *CamVid* test set, the PIDNet scored 82.0% mIoU. As for the *Cityscapes* test set, it was proved that PIDNet shows the best trade-off between inference speed and accuracy, achieving an mIoU of 80.1% (PIDNet-M) and 80.6% (PIDNet-L).

6.4 Dilated Convolutional Methods

Dilated Convolution or atrous Convolution method is a mechanism developed to increase the receptive field of the convolutional neural networks. They work by adding gaps between kernel elements within a convolution layer, which is controlled by the dilated (atrous) rate. This results in the model being able to cover a wider area while capturing contextual information and preserving fine features. Additionally, the spatial resolution of feature maps is maintained. However, models that utilize this technique can be more complex, requiring more memory consumption and being more difficult to train. Furthermore, they have a limited understanding of local context.

In scene parsing, objects can be encountered in various scenes, exhibiting multiple shapes and sizes. The model must possess the capability to classify them accurately. Up until the introduction of the pyramid scene parsing network or PSPNet by [88], FCNs lacked the ability to collect contextual information efficiently, leading to misclassification of object classes. To address this limitation, Zhao et al. developed a multi-scale network, comprising a residual network followed by a dilated network. The ResNet backbone is employed to extract features, which are then passed to a novel Pyramid Pooling Module. This module fuses feature maps from four different pyramid scales, which are subsequently upsampled and concatenated to create a comprehensive feature representation incorporating both local and global context information. The resulting features are then passed through a convolutional layer to obtain the final prediction. Initially designed for scene parsing, PSPNet later found applications in semantic segmentation. Its performance was evaluated on the *PASCAL VOC* dataset, achieving an 82.6% mIoU accuracy without pre-training on *MS-COCO* and 85.4%

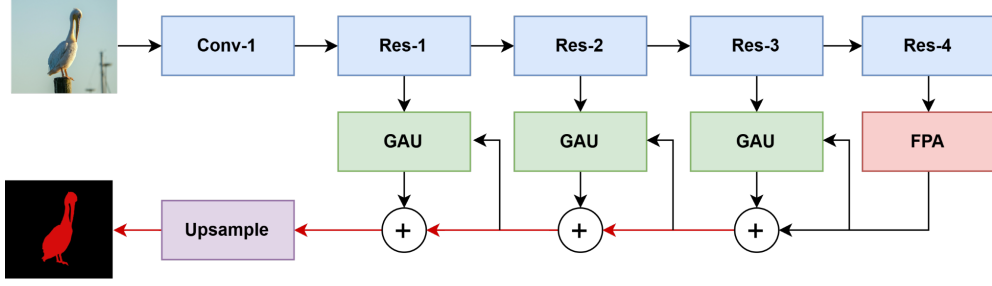


Fig. 10 Representation of Pyramid Attention Network showcasing the use of both of introduced modules FPA and GAU proposed by [91]

with pre-training. Additionally, on the *Cityscapes* test set, it achieved 80.2% accuracy using fine and coarse data.

[89] introduced atrous convolution into modules such as the Atrous Spatial Pyramid Pooling module (ASPP) to capture multi-scale context. The DeepLabv3 model was tested on the *PASCAL* dataset and achieved 85.7% accuracy. When tested with the implementation on the ResNet-101 model, it achieved a performance of 86.9%. The DeepLabv3 model was also tested on the *Cityscapes* test set and achieved an 81.3% mIoU. [90] improved upon the DeepLabv3 model with DeepLabv3+. By adding a new decoder module, they were able to refine object boundaries and achieve more precise results. The DeepLabv3+ model was tested on the *PASCAL VOC* dataset and reached 87.8% mIoU. Additionally, it was tested on the *Cityscapes* dataset and yielded an accuracy of 82.1% mIoU.

When using dilated convolutions like those in Deeplab’s ASPP module [89] or PSPNet’s pyramid pooling module [88], there may be a loss of local pixel information. To solve this problem, [91] created the Feature Pyramid Attention module (FPA), shown in Figure 10. This module combines context information from different pyramid scales and globally pools it to create a better feature representation. Li et al. also found that current decoder architectures lack diverse scales of low-level feature map information, resulting in lower-resolution output. To overcome this limitation, the authors introduced the Global Attention Upsample module (GAU) on each decoder. This module uses global context to guide the selection of category localization details. The Pyramid Attention Network (PAN) achieved 84.0% mIoU on the *PASCAL VOC* dataset and 78.6% mIoU on *Cityscapes* using ResNet-101 as a backbone without coarse annotations.

Many models for semantic segmentation use the ImageNet pre-trained as their backbone. However, this has a limited field of view, so a special contextual module is often added to address this issue. [92] proposes a different approach by creating a backbone specifically for semantic segmentation called RegSeg. By introducing a D-block, which is a dilated block structure, and maintaining a low number of channels, they can increase the field of view while retaining the local details of the image. The innovation lies in using group convolutions in the D-block and varying the dilation rates in each group to extract multi-scale features. Testing on the *CamVid* dataset

showed an accuracy of 80.9%, with 70 FPS. On the *Cityscapes* dataset, it achieved a performance of 78.13%, with 30 FPS.

6.5 Attention-Based Methods

Attention-based methods are frameworks that use attention mechanisms to help models determine the significance of different regions or features in an image. Strategies like feature (channel) attention, spatial attention and self-attention, which is mostly found in transformer-based models, are included on multiple stages of image segmentation. Attention-based models are known for their ability to handle complex scenes with occluded objects and structures, and for their better understanding of overall context. However, these methods can increase model complexity and require careful hyperparameter tuning, which can increase computational cost. Additionally, they require large amounts of data.

[93] introduced a new module that can be integrated with any backbone architecture to improve semantic segmentation by incorporating richer context. Their approach involves using class centers to extract global features from a categorical perspective, thereby achieving class-level features that represent all classes in an image. To accomplish this, they developed an Attention Class Feature module comprising a Class Center Block and a Class Attention Block that identify the features of the class center and the class attention, respectively, to create an attentional class feature map. Finally, the attentional class feature map and the feature map are combined to generate the final segmentation prediction. The ACFNet was evaluated on the *Cityscapes* dataset using ResNet-101 as the backbone architecture, and achieved a mIoU of 81.8%.

[94] proposed a new module called criss-cross attention integrated on the CCNet model to address the complexity of attention-based models. This module operates by taking a local feature map and applying three convolutional layers in parallel. The outputs of two layers are then combined using an affinity operation followed by a softmax layer to create an attention map. The third output is combined with the attention map through aggregation, improving the local features and pixel-wise representation. However, the final feature map only captures contextual information in horizontal and vertical directions, but not the connection between pixels and their surroundings. To address this, the feature map is fed through another criss-cross attention module to achieve denser context information, known as the recurrent criss-cross attention (RCCA) module. The CCNet architecture uses a CNN for feature extraction, followed by dilation convolutions to retain more details. The local feature map from the CNN is then fed through the RCCA. The local features outputted by the CNN are concatenated and passed through several convolutions with the dense contextual feature map from the RCCA module to obtain the final prediction. This approach was evaluated on the *Cityscapes* and *ADE20K* datasets, using ResNet as the backbone. The model achieved 81.4% mIoU on the *Cityscapes* and 45.2% mIoU on the *ADE20K* validation set. It was also tested on the *COCO* validation set for instance segmentation, achieving an average precision of 37.3%.

[95] addressed the challenge of context aggregation in semantic segmentation. They proposed a three-phase approach for describing a pixel based on the representation of its analogous object class. Firstly, supervised learning was used to teach the model to

identify object regions by using ground truth segmentation. Then, the values of the pixels within each region were added up and used to calculate the relationship between each pixel and object region. Finally, the object-contextual representation, which is a weighted aggregation of all object region representations, was used to improve each pixel’s representation. The authors tested their Object-Contextual Representation (OCR) method on various datasets, including *Cityscapes*, using HRNetV2-W48 [96] as the experimental backbone. Their test on *Cityscapes* yielded an 84.2% mIoU performance.

[97] address the issue of context aggregation in semantic segmentation. They propose a method to distinguish between intra-class and inter-class context by introducing a new layer called Context Prior. This layer embeds Affinity Loss to supervise the learning process. When combined with the backbone network, it creates the Context Prior Network (CPNet). After the backbone network extracts the features, an aggregation module is used to join the spatial information and determine the contextual link. The Context Prior Map is then generated and used by Affinity Loss to create an Ideal Affinity Map. Additionally, two feature maps are generated: the intra-prior and inter-prior. By applying transformations such as reshaping and matrix multiplication, the inter-class and intra-class Context are captured and fed to a convolution layer for per-pixel prediction. The CPNet was tested on two datasets: *ADE20K* and *Cityscapes*. By adopting the ResNet architecture, it achieved a mIoU of 46.27% on the *ADE20K* dataset and 81.3% on the *Cityscapes* dataset.

[76] developed a new pipeline for semantic segmentation that uses convolutional networks without transformers. Their method involved an encoder-decoder architecture with a pyramid structure for the encoder. Instead of a self-attention mechanism, the team uses a new module called the multi-scale convolutional attention (MSCA) module, which is made up of a depth-wise convolution, multi-branch depth-wise strip convolution, and a 1x1 convolution. The output of the 1x1 convolution serves as attention weights to reweight the input of MSCA. The team also developed another module called MSCAN, which is a sequence of these building blocks assembled. The decoder employs a lightweight Hamburger structure and just aggregates the features of the last three stages to extract the global context. The team evaluated their approach on multiple datasets such as *Cityscapes* and *ADE20K*. On *Cityscapes*, SegNeXt-T achieved 79.8% mIoU and 78.0% mIoU with 25 FPS. Other variations with more parameters were also evaluated, including SegNeXt-S, which scored 81.3% mIoU, SegNeXt-B, which obtained 82.6% mIoU, and SegNeXt-L, which reached 83.2% mIoU. On the *ADE20K* dataset, SegNeXt-L results were 51.0% mIoU.

6.6 Transformer Methods

Methods based on Transformers employ Transformer architectures to process images. At the core of these methods are self-attention layers and feedforward neural networks that collaborate to capture features and improve them. Although these methods are still complex, necessitating significant computational resources and enormous amounts of data [44], they hold great potential for the future of semantic segmentation because of their ability to achieve high accuracy in highly complex scenes.

There have been recent breakthroughs in the field of computer vision that have dismantled the CNN paradigm.

Classic encoder-decoder frameworks based on CNNs still struggle to capture long-range dependencies and incorporate global context, especially on complex images. Inspired by these issues, [98] proposed an encoder based on a transformer replacing the traditional stack of convolution layers. The SEgmentation TRansformer (SETR) sees semantic segmentation as a sequence-to-sequence prediction task. The transformer encoder is fed by a sequence of vectors resulting from the image split by fixed-size patches, which are then linearly embedded and added to position embedding. The transformer with global receptive fields in each layer learns the feature representations. The authors designed three different decoders to do pixel-level segmentation. The SETR-Naive projects the transformer features to the category dimension by utilizing a two-layer network (1x1 conv + batch norm + ReLU + 1x1 conv). Afterward, to get the full-resolution image, a bilinear upsampling method is employed. The SETR-PUP utilizes progressive upsampling, thus getting a less noisy final output that preserves spatial details. The SETR-MLA decoder aggregates features from multiple uniformly distributed layers of the transformer encoder. This strategy improves the integration of multi-level contextual information. The SETR-PUP was assessed on the *Cityscapes* test set, achieving 81.64 mIoU. On the *ADE20K* dataset, the SETR-MLA scored 50.28%.

[99] expanded the ViT [44] to create Segmenter. This framework is an encoder-decoder architecture based on a transformer. Initially, the image is divided into an array of patches, to which positional embeddings are added to capture spatial information. These embeddings undergo encoding by the Transformer encoder, producing an encoded patch with semantic information. The decoder then takes these embedded patches and converts them into class scores, which are further input to a mask transformer. The mask transformer utilizes a set of learnable class embeddings to obtain pixel-wise class scores, ultimately yielding the final prediction. The Segmenter’s performance was assessed on the *ADE20K* and *Cityscapes* validation sets. With ViT serving as the backbone, the model achieved a 51.8% mIoU on *ADE20K* and 81.3% on *Cityscapes*.

[100] realized that most Transformer-based semantic segmentation methods disregarded the decoder’s contributions to improve results. To address this, the authors developed a new model, which features novel encoder-decoder mechanisms. First, they introduced an encoder called MiT, which is capable of producing both high-resolution fine and low-resolution coarse features thanks to its hierarchical structure. Additionally, this encoder is adaptable to different resolutions as it avoids interpolating positional code. Next, they proposed a lightweight multi-layer perceptron encoder that leverages Transformer features to gather information both locally and globally from different layers and combine them to create precise segmentation masks. The Segformer model was tested on the *Cityscapes* and *ADE20K* datasets, achieving 84.0% mIoU and 51.8% using MiT-B5 as the backbone, respectively. Furthermore, using MiT-B0 on the *Cityscapes* test set, Segformer scored 71% mIoU with 47.6 FPS.

[101] developed an adaption of a vision-specific transformer (ViT) to try to solve some of the performance gaps that this architecture still holds when applied to computer vision. The innovation was to introduce a vision-specific inductive bias, thus avoiding modifying the original architecture. These biases are achieved by three modules: a spatial prior module for capturing the local semantics from input images, a spatial feature injector for incorporating spatial prior into the ViT and a multi-scale feature extractor to reconstruct the multi-scale features required by dense predictions tasks. This implementation was tested for the ViT-T/S/B models with the use of ImageNet-1K pre-training and with ImageNet-22K weights for the ViT-L model. On the *ADE20K* test set, ViT-adapter-L achieved 54.7 with multi-scale testing.

[102] developed the Lawin Transformer to improve ViT’s efficiency and reduce computation costs in the semantic segmentation task. This adaptation uses a window attention mechanism to create multi-scale representations. They designed a decoder that applies a large window attention, allowing the local window to query a wider context with minimal processing overhead. This approach captures contextual information at various scales. Moreover, they added an efficient hierarchical vision transformer (HVT) beneath the Lawin Transformer to incorporate multi-scale representations into the semantic segmentation vision transformer. The Lawin Transformer was tested on various datasets, including *Cityscapes* and *ADE20K*. When using Swin Transformer as a backbone on *Cityscapes*, the Lawin achieved 84.4% mIoU. On *ADE20K*, it scored 56.2% mIoU with the same backbone.

[103] developed a solution to address the high computation cost of Transformers in semantic segmentation. Their approach involves a new block called RTFormer, which consists of a low-resolution branch and a high-resolution branch. The low-resolution branch utilizes a GPU-Friendly Attention technique that integrates matrix multiplication to obtain high-level global context while being less taxing on the GPU. The high-resolution branch employs cross-resolution attention to apply the context information obtained from the low-resolution branch to each high-resolution pixel. Additionally, a stepped layout is used to provide more representative features from the low-resolution branch to the high-resolution branch. The RTFormer block is integrated into an RTFormer architecture that uses basic convolutional blocks to extract features, which are then fed to the RTFormer block. The low-resolution feature information passes through a segmentation head, a Deep Aggregation Pyramid Pooling Module (DAPPM) module, and finally through a classification head to make the prediction. The RTFormer was compared to other state-of-the-art models on datasets such as *Cityscapes* and *CamVid*. The RTFormer-Slim achieved 76.3% mIoU on *Cityscapes*, while the RTFormer-Base, which has a larger cross-feature spatial size, achieved 79.3% mIoU. On the *CamVid* test set, the RTFormer-Slim achieved 81.4% mIoU and the RTFormer-Base achieved 82.5%.

[104]’s research served as a foundation for the work of [105] that created the PolyMaX architecture. They aimed to design a simple model that could perform a range of dense prediction tasks, such as image segmentation, depth estimation, and surface normal estimation. To achieve this, they developed a mask transformer framework, which uses a cluster-based approach to prediction. Instead of classifying each pixel individually, similar pixels or instances are grouped into clusters and classified

together. This approach enabled the researchers to create two branches within the pipeline: an encoder/decoder for extracting pixel features and a transformer decoder for taking queries as inputs and associating them with pixel features, resulting in object queries or cluster centers. The per-pixel embeddings from the first branch are then multiplied by the cluster centers and passed through a softmax function to create a distribution map for each task. In the semantic segmentation task, the output, the maximum probability semantic label, is identified using the argmax operation, which is applied on top of the previous operations. The PolyMaX architecture was tested on the NYUD-v2 *NYUD-v2* [106] test set, achieving 58.08% mean Intersection over Union using ConvNeXt-L as the encoder.

7 Instance Segmentation

As previously discussed, instance segmentation is the process of segmenting items in an image independent of its class and type. It is useful for counting the number of objects and distinguishing their shapes, culminating in a mix of object detection with semantic segmentation. In this section, the methods will be grouped following [117] survey logic, which is essentially by their composition such as two-stage methods, multi-stage methods, and single-stage methods. Furthermore, Table 3 presents the discussed methods, followed by the average precision and speed.

7.1 Two-stage methods

Two-stage methods are based on the pipeline of positioning objects and generating masks, therefore they can be further divided into top-down methods or bottom-up methods. The two-stage methods are known for their success in instance segmentation, yet due to their sequential nature, if the object detection algorithm works less effectively, the ultimate result will suffer greatly.

7.1.1 Top-down methods

Top-down techniques work sequentially. First, the region proposal algorithm detects regions of interest, which are then fed to another model for instance segmentation.

The Fast R-CNN framework was designed to improve accuracy and speed up the process of training and testing, drawing on the foundational principles of the Region-based Convolutional Network (R-CNN) [118] method. [119] introduced an end-to-end detector training process by inputting the image and a set of regions of interest (RoI). Each region of interest (RoI) is first pooled into a feature map of a pre-defined size. Then, it is mapped by fully connected layers to a feature map. Finally, the output consists of two parallel branches: one for the softmax probability of category object prediction and the other for per-class proposal refinement offsets.

In their work, [120] introduced Faster R-CNN, a pipeline designed to improve the performance of Fast R-CNN. This was achieved through the integration of a module called Region Proposal Network (RPN) into the Fast R-CNN detector. RPN is a fully convolutional network that uses attention mechanisms to propose regions. These proposed regions are then used as input to the Fast R-CNN detector, resulting in

Table 2 Experimental results of semantic segmentation methods on the test set of *Cityscapes* (green), *CamVid* (orange) and *ADE20K* (red). ... designates information not provided.* refers to data not reported. ~ stands for results on the validation set of the datasets. The *R* designates the ResNet architecture. The results are presented in percentage. All relevant data is incorporated within the exploratory framework

Type	Name	Backbone	Parameters	mIoU	FPS	mIoU	FPS	PQ	FPS
Fully Convolutional Networks	FCN [67]	—	...	*	*	*	*	*	*
	—	—	...	*	*	*	*	*	*
Encoder-Decoder	U-Net [69]	—	...	*	*	*	*	*	*
	FASSD-Net [70]	—	2.85M	78.8	41.1	*	*	*	*
	PP-LiteSeg [71]	STDC2/STDC1 [84]	...	77.5/72.0	102.6/273.6	75.0/73.3	154.8/222.3	*	*
	MetaSeg [75]	MSCAN-B [76]	29.6M	82.7	*	*	*	48.5	*
	—	—	—	—	—	—	—	—	—
Multi-Scale and Pyramid Networks	ICNet [77]	PSPNet [88]	...	70.6	30.3	67.1	27.8	*	*
	—	R18/DF1 [37, 79]
	SFNet [78]	R18 [37]	50.3M/9M/12.9M	79.8/74.5	33.3/134.5	*73.8/70.4	*35.5/134.1	*44.7	*
	—	R101 [37]
	BiSeNet [82]	R18/Xception39 [37, 57]	49.0M/5.8M	74.7/68.4	65.5/105.8	68.7	*	*	*
	—	R18 [37]
	BiSeNetV2 [83]	—	...	75.3/72.6	47.3/156	78.5/72.4	32.7/124.5	*	*
	STDC [84]	STDC2/STDC1 [84]	12.5M/8.44M	76.8/71.9	97.0/250.4	73.9/73.0	152.2/197.6	*	*
	SFNet-Lite [80]	STDC2/STDC1 [84]	13.7M/9.7M	79.0/78.8	92.3/119.1	*	*	*	*
	BiSeNetV3 [85]	STDC2/STDC1 [84]	12.5M/8.44M	79.0/73.5	93.8/244.3	76.6/75.1	147.6/198.4	*	*
Dilated Convolutional	PIDNet [87]	PIDNet [87]	36.9M/7.6M	80.6/78.2	31.1/100.8	82.0/80.1	85.6/153.7	*	*
	—	—	—	—	—	—	—	—	—
	PSPNet [88]	R101 [37]	...	80.2	*	*	*	*	*
	DeepLabv3 [89]	R101 [37]	...	81.3	*	*	*	*	*
	DeepLabv3+ [90]	Xception-71 [57]	...	82.1	*	*	*	*	*
	PAN [91]	R101 [37]	...	78.6	*	*	*	*	*
Attention	RegSeg [92]	—	3.34M	79.1/77.5	11.3/18.3	80.9	112	*	*
	—	—	—	—	—	—	—	—	—
	ACFNet [93]	R101 [37]	...	81.8	*	*	*	*	*
	CCNet [94]	R101 [37]	...	81.4	*	*	*	45.2 ~	*
	OCR [95]	HRNetV2-W48 [96]	10.5M	84.2	*	*	*	45.7 ~	*
	CPNet [97]	R101 [37]	...	81.3	*	*	*	46.3 ~	*
Transformer	SegNeXt [76]	SegNext	48.9M/4.3M	78.0/83.2	25.0	*	*	51.0	*
	—	—	—	—	—	—	—	—	—
	SETR [98]	T-Large [98]	318.3M/310.6M	81.6	*	*	*	50.3	*
	Segmenter ~ [99]	ViT-L [44]	307M	81.3	*	*	*	51.8	*
	Segformer ~ [100]	MiT-B5/MiT-B0 [100]	84.7M/3.8M	84.0/71.9	2.5/47.6	*	*	51.8/37.4	9.8/50.5
	ViT-Adapter ~ [101]	BEiTv2 [107]	571M	*	*	*	*	61.2	*
	Lawin Transformer [102]	Swin-L [58]	201.2M	84.4	*	*	*	56.2	*
	RTFormer [103]	—	16.8M/4.8M	79.3/76.3 ~	39.1/110.0	82.5/81.4	94.0/190.7	42.1/36.7 ~	71.4/187.9
Panoptic	PolyMaX [105]	—	...	*	*	*	*	*	*
	—	—	—	—	—	—	—	—	—
	UPsNet [108]	R101 [37]	46.1M	79.2	*	*	*	*	*
	Panoptic-DeepLab [109]	Xception-71 [57]	46.7M	84.2	*	*	*	*	*
	Axial-DeepLab [110]	Axial-R-XL [110]	173M	84.1	*	*	*	*	*
	PanoNet ~ [111]	ICNet [77]	12M	74.6	20	*	*	*	*
	EfficientPS ~ [112]	—	40.9M	82.1	*	*	*	*	*
	K-net+UperNet ~ [81, 113]	Swin-L [58]	...	*	*	*	*	54.3	*
	MaskFormer ~ [114]	R101 R50 [37]	60M 41M	81.4	*	*	*	49.7	*
	CMT-DeepLab [115]	Axial-R50 [110]	95M	81.4	*	*	*	*	*
	Mask2Former ~ [114]	Swin-L [58]	216M	83.3	*	*	*	56.4	*
	kMAX-DeepLab ~ [104]	ConvNeXt-L [64]	232M	83.5	3.1	*	*	55.2	4.0
	MP-Former ~ [116]	Swin-L [58]	216M	83.9	*	*	*	56.9	*

improved object detection accuracy and the benefit of near real-time FPS during operation.

[61] improved the Faster R-CNN architecture by introducing a third branch, resulting in better object detection capabilities and the implementation of high-quality segmentation mask generation to the framework. Mask R-CNN was tested on the *COCO* with ResNeXt-101-FPN as the backbone, scoring an average precision percentage of 37.1%. It was also tested on the *Cityscapes* dataset, yielding 32.0% AP with ResNet-FPN-50 as the backbone.

[73] added a MaskIoU head block to the Mask R-CNN architecture. The mask Intersection over Union (IoU) is used to measure the pixel-wise overlap between the predicted mask and the ground truth mask for each predicted instance. If the mask IoU is higher than a certain threshold, it is considered a true positive, indicating that the model’s segmentation is accurate. Conversely, if the mask IoU falls below the threshold, it’s classified as a false positive. Precision and recall are calculated based on the true positives, false positives, and false negatives, and this information is used to create the precision-recall curve. This improvement prioritizes more accurate mask predictions, which in turn improves instance segmentation. When challenged on the *COCO* test set, the Mask Scoring R-CNN with ResNet-101-DCN-FPN backbone yielded a 39.6% AP.

[121] made further improvements to Mask R-CNN by addressing the issue of information propagation. To tackle this problem, they developed a pipeline that utilizes features at low levels to help segment instances. This is achieved by first introducing a bottom-up path augmentation to the feature pyramid network to improve the accuracy of the localization signals at low levels, thereby shortening the information path. Next, an adaptive feature pooling component is integrated to aggregate features from all feature levels and create simple paths. These paths are then fed to a small fully connected network, which branches out with similar outputs as Mask R-CNN, ultimately improving the mask quality. PANet achieved a score of 42.0% AP on the *COCO* test set using ResNeXt-101 as the backbone. Nevertheless, when it utilized an ensemble backbone comprising 3 ResNeXt-101 ($64 \times 4d$), 2 SE-ResNeXt-101 ($32 \times 4d$) [122], 1 ResNet-269 [123], and 1 SENet [122], it scored 46.7% AP. On *Cityscapes*, PANet used ResNet-50 as the backbone and achieved a score of 36.4% AP.

To improve instance segmentation performance, it can be helpful to refine segmentation boundaries through post-processing techniques. [124] developed a refinement method that they incorporated into their own method. The authors propose converting the masks of the segmentation network into polygons, allowing for better fitting of the deforming network on object boundaries. The PolyTransform method was tested on the *Cityscapes* reaching 40.1% AP. Additionally, this technique also demonstrates a 35% speed increase in annotating. Furthermore, it performs better than the boundary metric by 2.0% when compared to earlier work on annotation-in-the-loop.

[125] developed a pipeline for refined segmentation that first processes the input image through a boundary branch to obtain a binary boundary map. Then, using a direction branch, a direction map is predicted and added to the boundary map. The resulting output is then converted to an offset map and further refined to produce segmented results. The SegFix was added to several methods and tested on the *Cityscapes*

test set. The PANet method improved to 37.8% AP, and the PolyTransform scored AP of 41.2%.

7.1.2 Bottom-up methods

Bottom-up methods start with pixel-level or superpixel-level segmentation of the entire image, grouping it into segmented regions that are then classified.

[126] developed the MaskLab model, an extension of Faster-RCNN [120]. This model uses box predictions that are further refined by the box classifier, and then separates and identifies objects with semantic segmentation logits. It also employs direction prediction logits to differentiate between instances of the same semantic class. When challenged on the *COCO* dataset, with ResNet-101 as the backbone, it achieved an average precision of 38.1%.

The method of proposal-based segmentation involves detecting objects using a bounding box approach and then creating a binary mask. However, this method has low resolution and is ineffective for real-time applications. On the other hand, proposal-free instance segmentation has improved binary mask resolution and is faster due to the use of embedding loss functions, pixel affinity, and dense prediction networks. To maintain a high segmentation accuracy in real-time applications, a new clustering loss function has been proposed by [127] for proposal-free instance segmentation. This loss function groups together the spatial embeddings of pixels from the same instance, maximizing the intersection-over-union of the resulting instance mask while also learning an instance-specific clustering bandwidth. When combined with a quick architecture, this method enables real-time instance segmentation with high accuracy. The method was evaluated using the ERFNet [128] network architecture as a base network and tested on the *Cityscapes* instance segmentation challenge, achieving an AP-score of 27.6%.

[129] argue that proposal-free approaches do not take advantage of the relationship that can be extracted from semantic and affinity information. Therefore, the authors designed the SSAP framework, a single-shot proposal-free instance segmentation method, to fill that gap. In its first stage, it has a unified u-shape framework that learns the relationship between semantic information and affinity information, which then feeds the cascade graph partition that refines the instances masks outputting the final prediction. This method was evaluated on the *Cityscapes* test set, which scored 32.7% using only fine training. It was also tested on the *COCO* test set for panoptic segmentation, which yielded a panoptic quality percentage of 36.9, using ResNet-101 as the backbone.

[130] proposed a way to improve the boundaries of predicting instance masks by implementing a post-processing refinement framework to any model. Along the anticipated instance boundaries, the Boundary Patch Refinement (BPR) extracts a number of small boundary patches, assigned by the framework. Afterward, the image patches and mask patches go through the refinement network and are finally reassembled into a precise instance mask. The effectiveness of this framework was tested on *Cityscapes* using various models. The results showed that incorporating this framework into Mask R-CNN [61] led to an improvement in performance, achieving a score of 36.9 % AP.

Using SegFix [125] and BPR together, the PolyTransform [124] model achieved an average precision of 42.7%.

7.2 Multi-stage methods

Multi-stage methods take two-stage methods further by trying to fill the gap between the detection and segmentation tasks. Therefore, multi-stage methods rely on more fused information using cascade pipeline methods and attention methods which can achieve better performances and more precise instance masks. Nevertheless, the better performance comes with the drawback of higher computation requirements, and due to that it is harder to be applied in real-world applications.

7.2.1 Cascade-based

The cascade framework was first introduced by [131], where instance segmentation was divided into three tasks: determination of instances, where instances are represented by class-agnostic bounding boxes; mask definition, where pixel-wise masks are defined for each instance; and instance classification, giving a semantic category to each instance.

[132] designed a multi-stage architecture based on the Mask R-CNN framework for object detection. The authors proposed that stages that are found deeper on the cascade pipeline be more selective regarding close false positives. Additionally, the stages of the Cascade R-CNN are trained sequentially, where the output of one stage is the input of the other, decreasing the chance of overfitting. In 2019, [62] extended this framework to instance segmentation by adding an instance head to the cascade. The Cascade Mask-RCNN head was placed parallel to the detection branch where three possibilities were studied. By adding the segmentation branch in the first stage or the last stage, where in the last stage, it brings more examples, but can be harmful for object detection. Or adding a segmentation branch to each stage, which maximises the diversity of features to learn the segmentation mask prediction. The final model uses a three-stage cascade with a segmentation branch embedded on the first stage, which shows to bring the best AP vs computational cost trade-off. When evaluated on the *COCO* validation dataset, using ResNeXt-152 with enhancement techniques during training and inference, Cascade Mask-RCNN scored 42.3% AP.

The major problem with Cascade Mask R-CNN was that the processes of object detection and instance segmentation were done separately without sharing of information between them. To tackle this problem, [133] devised a hybrid framework called Hybrid Task Cascade, which permits the information to flow by adding connections between different mask branches and a better refinement of results by incorporating both cascade and multi-tasking at each stage. Furthermore, a fully convolutional network is adopted to enable the model to distinguish the background from clusters and, therefore, learn to discriminate features by providing spatial information. The HTC architecture was challenged in the *COCO* test set, using an ensemble of backbones including ResNeXt-101 [52] 64x4d and ResNeXt-101 32x8d, SENet-154 [122], DPN-107 [134] and FishNet [135], reaching 49.0% AP.

[136] presents a novel pipeline designed to address two fundamental challenges in computer vision. Firstly, the pipeline aims to accurately segment indistinguishable objects, particularly those that are partially occluded. Secondly, it attempts to differentiate instances through the utilization of relationships between different objects, such as their spatial proximity. To achieve this, the authors introduce an object mining strategy. In the initial stage, a feature pyramid network is employed to extract features, which are subsequently utilized by both a mask learning subnetwork and a semantic perceiving stage. The semantic perceiving stage identifies pixels that likely belong to distinct instances, denoted as original instance descriptors, which convey label and location information through a subnetwork based on semantic segmentation. Subsequently, an excavating method, integrated within the instance subnetwork, learns instances from the regions surrounding the original descriptors and employs them to identify potential indistinguishable instances, termed mined descriptors, which also encompass classification and spatial localization information. Both the original and mined descriptors are then incorporated into an instance purifying graph, wherein they are juxtaposed based on feature distances in a relationship graph, facilitating the comparison of instance similarity and the output of independent instances. These independent instances are further processed by the mask learning subnetwork to generate the final instance masks. The PEP (Perceiving, Excavating, and Purifying) pipeline was subjected to evaluation against state-of-the-art models on the *COCO* test set, yielding an average precision of 40.9% when employing ResNet-101 FPN as the backbone architecture.

7.2.2 Attention-based

Attention-based methods are pipelines that employ multiple layers or steps of attention mechanisms inside a neural network architecture to boost the prediction of spatial connections and contextual information throughout the image. These techniques use attention processes to focus on different features of the input data at different levels of the process, resulting in more robust and accurate instance segmentation results.

A new approach to segmentation has emerged with the development of self-attention methods that rely on query-based pipelines. This method employs bipartite matching loss and learning query embeddings to match sparse ground truth objects.

[137] use parallel supervision on dynamic heads to perform instance segmentation. QueryInst augments the Sparse R-CNN [138] with a mask pooling operator P^{mask} and a box dynamic convolution module $DynConv^{mask}$. Similar to the pooling operator P^{bbox} , P^{mask} uses a standard RoI-Pooler, such as RoI-Align, to extract the characteristics of the current stage instance mask. The $DynConv^{mask}$ block is then created to connect the relationship between the query embedding q_{t-1} of the current stage and the instance mask features. Particularly, the $DynConv^{mask}$ block improves the instance mask by adding two successive convolutional layers to the x_t mask, whose kernel parameters are generated by the query embedding q_{t-1} . The next stage, which is repeated numerous times to improve performance, receives as input the object box prediction from the previous stage. When evaluated on the *COCO* test set with 300 queries using Swin as the backbone it achieved 49.1% AP.

[139] present an end-to-end object detector based on three parts. The first part uses ResNet as a backbone for feature extraction that is fed to a transformer encoder composed of a multi-head self-attention and a feedforward network to enhance the feature maps. The transformer decoder then receives the learnable object queries and generates the instance-aware query embeddings, which represent the features of each instance. These embeddings serve as input to the Unified Query Representation, which performs three tasks: classification, localization, and segmentation. The classification branch uses a fully connected layer, while the localization and segmentation branches use a multi-layer perception with a different number of hidden layers. During training, the predicted mask vectors are compared to the ground truth mask vectors generated by the spatial mask. The predicted mask is then reconstructed through the inverse process of compression coding. The SOLQ was tested on the *COCO* dataset, using Swin as the backbone, and achieved an AP of 46.7%.

[140] conducted a study to determine the effectiveness of query-based models in solving real-world problems. Their model consisted of three modules - backbone, pixel decoder, and Transformer decoder - where the first two extracted and refined multi-scale feature maps. The FastInst model, based on the Mask2Former [114] model, introduced three innovations in the Transformer decoder to achieve real-time segmentation. To improve the quality of the Transformer decoder’s embedding information and reduce the iteration update, the authors incorporated dynamic instance activation-guided queries. This process selected pixel embeddings with high semantics from the feature maps as initial queries. The Transformer decoder adopts a dual-path architecture, alternately updating the query and pixel features, resulting in more refined and fine-grained feature embeddings without the need for heavy pixel decoders. The model used the fine-grained feature embeddings to predict object class and segmentation masks at each layer. During the Transformer decoder training process, a ground truth mask-guided learning was introduced to standard masked attention to avoid a suboptimal query update process. This ensured that every query saw the predicted object in its entirety during training. The FastInst model’s performance was evaluated on the *COCO* dataset, using ResNet-50 as the backbone. The FastInst-D1 model was the fastest, achieving a frame per second value of 53.8 while achieving 38.6% AP. On the other hand, the FastInst-D3 model, utilizing the ResNet-50-d-DCN, achieved 40.5% AP while performing at 32.5 FPS.

7.3 Single-stage methods

Motivated to tackle the problems that two-stage and multiple-stage methods have, the single-stage methods were proposed. Using a single architecture to do both object detection and segmentation while trying to achieve real-time performance were the goals of this method. Inspired by the single-stage object detection algorithms, this method can be divided into anchor-based methods and anchor-free methods. Besides, there is a section about Transformers, since this method is rising in the computer vision world.

7.3.1 Anchor-based methods

Anchor-based methods utilize pre-trained detector frameworks as their backbone to generate anchors or predefined bounding boxes in an image. These anchors serve as a reference for object localization, classification, and prediction of instance masks. Thanks to their detector background, these methods are both fast and accurate. Nevertheless, to achieve optimal performance, these models require a significant amount of labelled data and careful tuning of their hyperparameters.

[141] designed the TensorMask framework, which utilizes dense sliding-window instance segmentation. The model employs 4D tensors to represent feature masks, with two dimensions representing object position and the other two representing relative mask position. To capture both large objects with high-resolution masks and small objects with low-resolution masks, the model uses a novel tensor bipyramid, similar to a feature map pyramid, but with two pyramids, one of which is reversed. The model’s performance was evaluated on the *COCO* test set, utilizing ResNet-101-FPN as the backbone, and achieved a score of 37.1% AP.

In an effort to improve real-time application results, [142] developed YOLACT, a fully convolutional model that splits instance segmentation into two parallel subtasks. The model generates prototype masks and predicts per-instance mask coefficients that combine to produce instance masks. This approach boosts runtime speed while maintaining coherence in the feature space. The team tested the model on the *COCO* dataset using ResNet-101-FPN as the backbone and achieved an AP of 29.8 while running at 33.5 FPS. Nevertheless, YOLACT falls short compared to other state-of-the-art models. To enhance this strategy [34], integrated deformable convolutions into the network ResNet-101, added a novel fast mask re-scoring, and optimized the prediction head with improved anchor scales and aspect ratios. Testing YOLACT++ on the same dataset revealed a performance improvement, with an AP of 34.1% achieved while maintaining the same FPS.

[143] proposed a framework for single-shot instance segmentation in their paper. The framework uses a mask prediction method that creates a compact vector of extracted information from the original mask. This method is applied on an additional branch of the MEInst framework for mask regression, which outputs the instance masks that are reconstructed later. The performance of MEInst was tested on the *COCO* dataset. The ResNeXt-101-FPN-DCN backbone was used and MEInst achieved an AP of 38.2%.

7.3.2 Anchor-free methods

Anchor-free methods aim to detect objects and segment their instances directly from the image without the use of anchors. Instead, they use keypoint detection to identify particular points such as an object’s corners or center. By using these points, the method is able to segment instances based on their interrelations. These methods are often precise due to their methodology of focusing on distinctive points. Furthermore, it is more straightforward than anchor-based methods and can be employed at any point during object detection. However, it may necessitate a substantial amount of

labelled data and may not be as effective in complicated scenes with obscured objects, which may require alternative strategies.

The primary objective of the instance segmentation task is to generate instance masks. Most algorithms follow either the "detect-then-segment" approach or first predict embedding vectors and then apply clustering techniques to group pixels into separate instances. [144] introduced the concept of "instance categories". By utilizing the object's size and location, each pixel within the object is assigned a category, transforming instance segmentation into a one-shot classification problem. This technique is based on two fundamental principles: category prediction and instance mask generation. The input is divided into a uniform grid, and if the object center falls on a cell, this cell predicts the semantic category and segmentation of the object mask. SOLO, using the Resnet-101 architecture as its backbone, was tested on the *COCO* dataset, resulting in a score of 37.8% AP.

[145] developed a new framework introducing the concept of dynamic instance segmentation, building on their work on SOLO [144]. Instead of using all masks outputted from the system, a final mask containing the instance whose center is at a specific location is used, reducing the computational and memory effort. This is achieved by multiplying one level of the pyramid features with a convolutional kernel. The new method, called SOLOv2, was tested on the *COCO* dataset and achieved better results (39.7% AP) than SOLO (37.8% AP) using the same backbone architecture. Yet, using Resnet-DCN-101-FPN, SOLOv2 scored an average precision of 41.7%. Additionally, by adding a semantic segmentation branch analogous to the mask feature branch, SOLOv2 can perform Panoptic Segmentation. Using the same strategy to combine instance and semantic results as in Panoptic-FPN [74] and employing ResNet50-FPN as a backbone, SOLOv2 yielded 42.1% PQ on the *COCO* dataset.

When it comes to real-life applications, computational cost is a crucial factor to consider. Traditional instance segmentation methods are generally slower than object detection. However, [146] have addressed this problem by developing a new shape signature system called "Inner-center Radius" (IR). This system takes the inner-center of an object segment and transforms the contour into polar coordinates. Chebyshev polynomials are then utilized to create a coefficient vector, which is used as the shape descriptor and regressed by the network. In addition, a pipeline is presented that explicitly decodes the multiple object shapes with tensor operations such as multiplication and addition. The ESE-Seg has been tested on the *COCO* datasets using YOLOv3 [147] as the base detector, achieving 21.6 % AP on *COCO* with a time of 26.0ms.

When dealing with instance segmentation, color information is typically essential. According to [148] instance segmentation can be achieved by merging images and disparity information to create object masks. This is done by utilizing stereo cameras to determine geometric estimations and introducing disparity information through the Region of Interest (ROI) method. The GAIS-Net model was tested on the Cityscapes dataset, where it achieved 32.5% AP with only fine annotations and 37.1% AP with fine annotations and pre-training with the *COCO* dataset.

[149] proposed a framework to achieve real-time instance segmentation, introducing FASSST (Fast Attention-based Single-Stage Segmentation NeT) as a single-stage

method employing an attention mechanism to identify instances. The authors first introduced a new attention model called the instance attention model (IAM). This module receives raw features from the MobileNet-54-V2 head and directly learns class probabilities and instance coordinates through single-stage regression. The instance information then passes through a series of pyramid layers to extract regions of interest, followed by ROI feature fusion to merge them, facilitating the gathering of global and local context information. Subsequently, these undergo several smaller-sized convolutional layers to obtain the instance masks. FASSST was evaluated on both the *COCO* and *Cityscapes* datasets, achieving a 34.2% AP with 59.2 FPS on the *COCO* test set and an AP of 31.1% with 47.5 FPS on the *Cityscapes* dataset.

7.3.3 Transformer-based

Instance segmentation using transformer-based methods is becoming increasingly popular in computer vision. These methods are based on transformer architectures, originally used in natural language processing. They offer advantages such as capturing global context and focusing on fine features to improve instance mask predictions. However, they also have high computational costs and require careful configuration, as well as large amounts of training data [44].

[150] introduced an end-to-end pipeline that utilizes a new instance segmentation Transformer. This transformer is designed to predict a group of classes, bounding boxes, and mask embeddings for each instance. The framework consists of four steps, starting with the extraction of instance features using a CNN backbone with FPN. Then, a transformer with dynamic attention is used to learn the relations between instances and match the low-dimensional mask embeddings with the ground truth mask embedding to set the loss. Next, a set of prediction heads are integrated to detect and segment instances, along with an N-step recurrent update to improve the set of predictions. During testing on the *COCO* test set, the ISTR with the ResNet101-FPN backbone scored 39.9% AP while maintaining good run-time performance. [151] aimed to make significant advancements in the transformer-based model pipeline by addressing a fundamental limitation regarding the effective utilization of spatial information and mask embeddings. The authors revisited their framework and introduced several mechanisms to enhance their predictions. Specifically, the authors incorporated Mask Meta-Embeddings (MME) to address mask encoding-decoding as a mutual information maximization problem, resulting in a meta-formulation for various techniques. In addition to the transformer with dynamic attention, known as the Dynamic Box Predictor, a new module called the Mask Information Generator (MIG) was introduced to predict mask embeddings. The predicted mask embeddings were then used by the Mask Meta-Decoder (MMD) to generate the set of predictions. Moreover, the authors extended the MMD by introducing the Spatial Mask Tuner (SMT), which takes both embedding and spatial information to produce the masks. The ISTR was evaluated on the *COCO* and *Cityscapes* datasets. When using Swin-L as the backbone and the SMT on the *COCO* test set, it achieved 49.7% AP with 2.9 FPS. For the *Cityscapes* test set, it yielded an AP of 36.2%.

A team of researchers from Meta AI [25] aimed at creating the first foundation model for image segmentation. The project introduces a novel task called promptable

segmentation, inspired by NLP foundation models, that generates a valid segmentation mask when provided with any segmentation prompt, like "red flower". The model, named Segment Anything Model (SAM), consists of three components: an image encoder based on the MAE pre-trained Vision Transformer architecture, a flexible prompt encoder that considers both sparse and dense prompts, and a fast mask decoder. SAM underwent training on the Segment Anything Data Engine, which leverages interactive and automatic segmentation approaches to improve object recognition. This resulted in the creation of 11 million images with 1.1 billion dependable segmentation masks, all compiled in the SA-1B dataset.

[152] have developed an iteration of SAM called FastSAM, which significantly improves runtime speed while performing instance segmentation using a two-stage approach. The first stage, dubbed All-instance Segmentation (AIS), employs the YOLOv8 [153] architecture to produce instance segmentation. Following the segmentation of all objects and regions, FastSAM uses various prompt types to identify objects of interest. FastSAM achieved a zero-shot instance segmentation of 37.9% AP on the *COCO* dataset using ViTDet [154] bounding box as a prompt. While this two-stage approach increases runtime speed by over 50 times, further improvements are necessary to achieve comparable results to SAM on zero-shot instance segmentation.

8 Panoptic Segmentation

The following section highlights the main achievements in panoptic segmentation up to 2024. The structural layout of this section draws inspiration from the survey paper by [160]. It comprises four subsections: Top-down methods, Bottom-up methods, Single-path methods, and Transformer-based methods. Concluding this section, Table 4 presents the results of the methods discussed.

8.1 Top-down methods

Top-down methods in panoptic segmentation are similar to those used in instance segmentation, with both depending on a sequential mechanism. The top-down approach in panoptic segmentation uses object detection algorithms to locate and identify instances. Proposals are first employed to reduce the search space, and then bounding boxes or spatial representations are predicted. Then, a semantic segmentation model is used in the totality of the image to segment stuff. These segments are then merged with the proposals to create a panoptic segmentation mask. These methods usually perform well, holding high PQ scores. Nevertheless, due to their high precision, they rely heavily on computation processing, which makes them unsuitable for real-time applications. It is noteworthy that several strategies were devised to lower computational costs, namely the one-stage top-down methods, which instead of generating proposals, use other mechanisms to extract bounding boxes.

In this section, the only one-stage method presented is SpatialFlow [161]. Therefore, no distinction will be made.

[108] developed a new method to tackle the panoptic segmentation task. The implemented model utilizes Mask R-CNN [61] as the shared feature extraction mechanism that feeds both the instance and semantic segmentation heads simultaneously. The

Table 3 Experimental results of instance segmentation methods on the *COCO* (green) and *Cityscapes* (orange) test-dev. . . . designates data not provided. * refers to data not reported. ~ stands for results on the validation set. The *R* designates the ResNet architecture, while the *X* symbolises the ResNeXt architecture. The results are presented in percentage. All relevant data is incorporated within the exploratory framework

Type	Name	Backbone	Parameters	AP	FPS	AP	FPS
Top-Down	Fast R-CNN [119]	VGG [40]	...	*	*	*	*
	Faster R-CNN [120]	VGG [40]	28000	*	*	*	*
	Mask R-CNN [61]	X101-FPN [52, 155] R50-FPN [37, 155] Ensemble	...	37.1	5.0	32.0	*
	PANet [121]	R50 [37]	...	46.7	*	36.4	*
	Mask Scoring R-CNN [73]	R101-DCN-FPN [36, 37, 155]	...	39.6	*	*	*
	PolyTransform [124]	R50 [37]	...	*	*	40.1	*
	PANet +SegFix [125]	-----	...	*	*	37.8	*
	PolyTransform+SegFix [125]	-----	...	*	*	41.2	*
Bottom-up	MaskLab [126]	R101 (JFT) [37]	...	38.1	*	*	*
	[Neven et al., 2019] [127]	ERFNet	...	27.6	*	*	*
	SSAP [129]	R50 [37]	...	32.7	*	*	*
	Mask R-CNN +BPR [130]	-----	...	39.2	*	36.9	*
Cascade	Cascade Mask-RCNN ~ [62]	X152 [52]	...	42.3	*	*	*
	HTC [133]	X101-FPN [52, 155]	...	47.1	2.1	*	*
	PEP [136]	R101-FPN [37, 155]	...	40.9	*	*	*
		Swin-L					
Attention	QueryInst [137]	[58] R50 [37]	...	49.1	3.3	34.4	*
	SOLQ [139]	Swin-L [58]	...	46.7	*	*	*
	FastInst [140]	R50-DCN-D [156]	...	40.5	32.5	*	*
Anchor	TensorMask [141]	R101-FPN [37, 155]	...	37.1	*	*	*
	YOLACT [142]	R101-FPN [37, 155]	...	29.8	33.5	*	*
	MEInst [143]	X101-FPN-DCN [36, 52, 155]	...	38.2	*	*	*
	YOLACT++ [34]	R101-FPN [37, 155]	...	34.6	27.3	*	*
Anchor-free	ESE-Seg ~ [146]	YOLOv3-Cheby [146]	...	21.6	*	*	*
	SOLO [144]	R101-FPN [37, 155]	...	37.8	10.4	*	*
	SOLOv2 [145]	R101-FPN-DCN [36, 52, 155]	...	41.7/ 37.1	*/31.3	*	*
	GAIS-Net [148]	R50-FPN [37, 155]	...	*	*	37.1	*
	FASSST [149]	MobileNet-54-V2 [54]	...	34.2	59.2	31.1	47.5
Transformer	ISTR 2021 [150]	R101-FPN [37, 155]	...	39.9	11.0	*	*
	SAM [25]	-----	...	46.5	*	*	*
	FastSAM [152]	-----	68M	37.9	*	*	*
	ISTR 2024 [151]	Swin-L [58] R50 [37]	...	49.7	2.9	36.2	*
Panoptic	UPSNet [108]	R101 [37]	46.1M	*	*	39.0	*
	Panoptic-DeepLab [109]	Xception-71 [57]	46.7M	39.0	*	*	*
	Axial-DeepLab [110]	Axial-ResNet-XL [110]	173M	*	*	39.6	*
	PanoNet ~ [111]	ICNet [77]	12M	*	*	23.1	20
	EfficientPS ~ [112]	-----	40.9M	*	*	43.8	*
	K-net [81]	R101-FPN [37, 155]	...	40.6	15.5	*	*
	Panoptic SegFormer [157]	R50 [37]	...	41.7	*	*	*
	Mask2Former ~ [114]	Swin-L [58]	216M	50.1	4.0	43.7	*
	kMAX-DeepLab ~ [104]	ConvNeXt-L [64]	232M	*	*	44.0	3.1
	YOSO ~ [158]	R50 [37]	...	35.6/33.0	33.5/46.1	*	*
	Mask DINO	Swin-L					

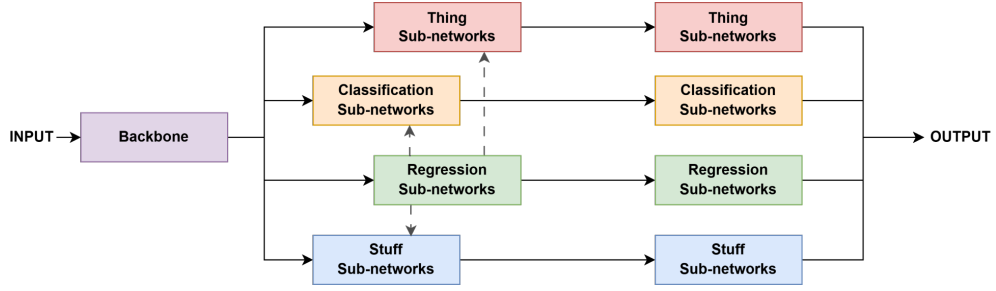


Fig. 11 Representation of Spatialflow network exhibiting the spatial flow mechanism between the subnetworks with dotted arrows as proposed by [162]

instance segmentation head outputs instance-aware representations as a bounding box, class, and mask. Simultaneously, the semantic segmentation head uses deformable convolution and multi-scale feature concatenation to output semantic segmentation. An upgrade that [108] did to the semantic head was adding Region of Interest (ROI) loss to emphasize foreground objects, which demonstrates improved performance. The panoptic head uses instance and semantic heads to classify pixels without needing parameters. The model also includes a classification feature to improve the identification of unknown classes by utilizing backpropagation from various modules. The performance of UPSNet was evaluated on both *COCO* and *Cityscapes* datasets. When tested in the *COCO* dataset, UPSNet used a ResNet-50-FPN as the backbone and reached 42.5% PQ, 78.0% SQ, and 52.4% RQ. In the *Cityscapes* dataset, UPSNet using ResNet-50 as backbone got 59.3% PQ, 79.7% SQ, and 73.0% RQ.

Chen et al. based their model on the RetinaNet [162] one-stage detector. The idea behind this decision arises from the fact that RetinaNet solely provides spatial context through its pixel-level features, which are subsequently utilized for both semantic and instance segmentation. Additionally, the authors project four parallel subnetworks to ensure feature fusing among tasks and create a spatial information flow that connects the four subnetworks. These networks are then connected to the four heads, as depicted in Figure 11. The thing head predicts instance masks, and the classification and regression head outputs together a bounding box to the thing head and the stuff head that gives semantic masks. SpatialFlow underwent testing on two datasets, namely *COCO* and *Cityscapes*. When ResNet-101-FPN served as the backbone on the *COCO* test set, it delivered 47.9% PQ, 81.7% SQ, and 57.6% RQ. On the *Cityscapes* validation set, it provided a PQ of 62.5%.

The fusion of instance and semantic masks presents a significant obstacle in achieving panoptic segmentation. The nature of instance masks is often incompatible with the semantic masks' results, making it harder to create an overall panoptic mask. Furthermore, if multiple instances are not correctly merged into a single mask, it can result in missing objects or inaccurate masks. Therefore, [163] designed two novel modules to solve this challenge. They have refined their tasks by improving their feature representation and creating more precise masks to address the issue of instance and semantic mask incompatibility. The approach they used involves the Iterative

Bi-directional Prediction Projection module, which serves to connect both tasks and allow for cross-task predictions via backpropagation paths. The Inter-instance Soft Occlusion Estimation module uses the bounding box trust ratings to refine the inter-instance occlusion output. The REFINE network, which was evaluated using the *COCO* dataset and ResNeXt-101-DCN [52] as its backbone, resulted in a PQ score of 51.1%, an SQ score of 82.6%, and an RQ score of 61.3%.

[164] introduce Ada-Segment, a novel approach that centers on improving the loss weight search mechanism as opposed to devising a new network architecture. This framework’s central point is the weight controller, which permits an adjustable training weight loss according to conditions by automatically producing new weights. The Ada-Segment pipeline consists of a backbone network for feature extraction, the Cascade R-CNN [62] as the detection branch, the SemanticFPN [74] as the semantic branch, and an end-to-end weight controller. The efficacy of this framework was evaluated on the *COCO* test set with the use of the ResNet-101-FPN-D backbone. It is worth noting that the "D" in the backbone acronym refers to deformable convolutions. The framework achieved 48.5% PQ, 81.8% SQ, and 58.2% RQ.

[112] introduce an unprecedented architecture called Efficient Panoptic Segmentation (EfficientPS). This innovative design incorporates its own novel shared backbone, EfficientNet-B5, which effectively encodes and integrates semantically rich multi-scale features. This architecture uses a variant of Mask R-CNN [61] as the instance head, a new proposed semantic segmentation head. Additionally, a novel panoptic fusion module is used to fuse the outputs of both heads to give a panoptic output. The researcher compared the method to state-of-the-art models using the *Cityscapes* dataset, training only on fine annotations and evaluating using a single scale. Various metrics were tested, including panoptic quality (PQ), segmentation quality (SQ), and recognition quality (RQ). The results showed a PQ of 64.1%, SQ of 82.6%, and RQ of 76.8%. Pre-training the model with *Mapillary Vistas* dataset resulted in a PQ of 67.1%, SQ of 83.4%, and RQ of 79.6%. [112] have developed the *KITTI* dataset, explicitly designed for panoptic segmentation. Additionally, they have conducted model testing on the dataset where the model scored 42.9% PQ on a single-scale evaluation and 43.7% PQ in a multi-scale evaluation

8.2 Bottom-up methods

Bottom-up methods result from directly employing a segment model, producing a unified framework. The segmentation model produces pixel-wise class labels and instance-specific segmentation masks by processing the entire image instantly. Therefore, although the process can achieve real-time performance, it sacrifices accuracy.

DeeperLab is a single-shot pipeline designed by [165] that incorporates several novel self-developed methods. The authors discuss the approach of using a fully convolutional network with a shared decoder and a single output for two-layer head predictions. They incorporate depthwise separable convolution and utilize mechanisms such as depth-to-space and space-to-depth instead of upsampling while increasing the kernel size. DeeperLab’s performance was put to the test on various datasets, including *Mapillary Vistas* and *COCO* datasets. In the *Mapillary Vistas* validation set, DeeperLab scored 32.0 % PQ with the Xception-71 backbone. Likewise, on the *COCO* test

set, DeeperLab, using the same backbone, achieved a PQ score of 34.3%, an SQ score of 77.1%, and an RQ score of 43.1%.

A new bottom-up system for panoptic segmentation is proposed by [109]. The goal is to accomplish comparable results with top-down approaches while increasing the inference speed. The Panoptic-DeepLab has the particularity of having a dual-ASPP, commonly used on semantic segmentation, and a dual-decoder used on instance segmentation. While the first branch possesses a regular design, the instance branch holds a class-agnostic detector with an instance center regression. Something deserving of mentioning is that this architecture during training only uses three loss functions and adds marginal parameters to semantic segmentation. The model was tested on *Cityscapes*, *Mapillary Vistas*, and *COCO* test sets. On the *Cityscapes* test set, Panoptic-Deeplab achieved a performance of 65.5% PQ. As for the *Mapillary Vistas* test set, it accomplished a PQ of 42.7%. Finally, on the *COCO* test set, it performed comparably well with other top-down methods, which use heavier backbones or deformable convolution.

[110] developed two distinct frameworks utilizing their self-attention method to perform panoptic segmentation. They introduced a new block consisting of an axial attention strategy that incorporates a position-sensitive method for precise positioning information over long distances. This ensured global connections while maintaining computational efficiency. This mechanism is integrated on the width and height axis, displayed sequentially. Through the use of this block, the complexity of the framework is diminished. They applied this block to ResNet, creating a new pipeline called Axial-ResNet, where two multi-head axial attention layers replaced the 3x3 convolution layer in the residual bottleneck block. DeepLab was adapted with the same changes as ResNet. With the additional removal of the final stride of feature extraction and the atrous spatial pyramid pooling module. When evaluated on the *COCO* and *Cityscapes*, the Axial-DeepLab-L which uses Axial-ResNet-L as the backbone, yields 44.2% PQ on *COCO* test set. On the *Cityscapes* test set, the Axial-DeepLab-XL reached 66.6% PQ.

PanoNet is a model that combines position information and feature masks for segmentation. The model uses ICNet [77], a lightweight network for semantic segmentation, as the backbone. While the semantic segmentation branch remains the same, [111] added a supplementary branch that uses pixel-level embedded maps with extra coordinate input channels. The mean-shift clustering algorithm then processes the map and generates instance masks that are combined with the semantic feature masks to produce the final output. After being tested on the *Cityscapes* dataset, PanoNet achieved a PQ of 55.1%, SQ of 77.5%, and RQ of 67.5%. It is worth noting that PanoNet can produce these results at a rapid pace of 20 FPS, indicating its ability to process high-quality images with minimal computational consumption.

8.3 Single-path methods

The single-path methods perform instance and semantic segmentation simultaneously, directly outputting panoptic segmentation masks. These methods have a simpler, compact approach that enables end-to-end learning, which simplifies training and improves computational efficiency.

Nonetheless, the combination of instance and semantic segmentation in the architecture design can result in complexity. Furthermore, challenges may arise when instances are in close proximity to one another, potentially compromising the accuracy of the model.

[166] developed a framework that jumps over the several steps that other methods need to do panoptic segmentation. With the introduction of a mask transformer, MaX-DeepLab is able to output directly segmentation masks, hence ending the need to use bounding boxes or object centers. The mask transformer is designed with a dual-path architecture, in which one branch carries a CNN and the other a transformer in constant communication, enabling the CNN to store memory in any layer. Furthermore, it introduces a novel metric for model optimization, which encompasses the product of the multiplication between class similarity and mask similarity. The training involves one-to-one bipartite matching to maximize similarity between ground truth and predicted masks. The MaX-DeepLab was challenged on the *COCO* dataset where it achieved 51.3% PQ.

[81] embraced the challenge of designing a pipeline that unified the three main segmentation tasks: instance, semantic, and panoptic. The framework is cemented in kernels, more precisely, dynamic kernels. First, a set of convolutional kernels equipped to learn about instance and semantic segmentation is randomly initialized. The union of the two provides panoptic masks. These kernels have a content-aware mechanism in order to be adaptive to changes in object location in the image and ensure a discerning ability to have better segmentation mask results. The bipartite matching mechanism was employed to process multiple instances in an image precisely, assigning learning targets to each kernel. Additionally, this framework only utilizes masks for learning. The K-Net method was tested on *COCO* and *ADE20K* datasets. In the former, using a CNN network, i.e., ResNet-101-FPN-DCN as the backbone, it achieved 48.3% PQ, while using a transformer-based backbone, i.e., Swin-L reached up to 55.2% PQ. In the latter, when testing only the semantic segmentation task, a K-Net+UperNet using Swin-L as the backbone achieved a mIoU of 54.3%.

[114] designed a mask classification model capable of consistently performing both instance and semantic segmentation. The framework uses three modules: a pixel-level module, a transformer module, and a segment module. The pixel-level module generates binary mask predictions by employing a pixel decoder that uses per-pixel embedding extracted from the feature maps. The transformer module generates class predictions using a set of transformer decoder layers that process per-segment embedding. Subsequently, the segment module, which obtains data from the remaining modules, infers the predictions. MaskFormer was tested on the *COCO* dataset, where it scored 53.3% PQ, 82.0% SQ, and 64.1% RQ.

[167] dug further into the K-Net to improve real-time applicability and performance. They developed a mask normalization strategy to ensure that the values are in the half-precision range and solve the problem of numeric overflow, increasing the inference speed. Furthermore, a mask pasting method is introduced that optimizes the post-processing step by using element-wise multiplication of masks and predicted classes. Instance-aware crop augmentation and instance discrimination loss are added to enhance the performance of instance kernels. Moreover, the backbone and neck of

the framework are updated to the novel RTFormer while including panoptic kernels right in the initialization stage. The *Cityscapes* and *ADE20K* datasets were used to test this solution. In the *Cityscapes* validation set, the RT-K-Net achieved 60.2% PQ, and in the *ADE20K* validation set achieved 33.2% PQ.

In a recent study, [158] aimed to develop a lightweight architecture for achieving real-time panoptic segmentation with their model, YOSO (You Only Segment Once). To accomplish this, they utilized a feature pyramid aggregator and introduced a convolution-first aggregation (CFA) module to merge multi-level feature maps. Additionally, they implemented a separable dynamic decoder to generate panoptic kernels and perform convolution for image feature maps. To perform multi-head cross-attention in a weight-sharing way, they used a separable dynamic convolution attention (SDCA) module on the decoder. The YOSO architecture was evaluated on various datasets, including the *COCO* and *Cityscapes* validation sets. YOSO achieved a PQ of 48.4% with 23.6 FPS on the *COCO* validation set using only ResNet-50 as a backbone. Even when scaling the input image to (512×800) , YOSO maintained a PQ of 46.4% with a FPS of 45.6. On the *Cityscapes* validation set, YOSO achieved a PQ of 59.7% and 11.1 FPS using the same backbone and larger image scale. When downscaling the image input, it reached 22.6 FPS while having a PQ of 52.5%.

[168] developed MaskConver, a fully convolutional architecture that demonstrates the continued relevance of convolutional methods. They propose a unified method for panoptic segmentation. Using their novel pixel-decoder, ConvNeXt-UNet, which consists of ConvNext [64] blocks disposed of similarly as U-Net [69] only in a higher level of the backbone (ResNet), which allows the capture of long-range context and high-level semantics in a more efficient way. Furthermore, three prediction heads that adopt depthwise convolution with a GeLU activation and large kernel 7x7 in order to reduce the computational cost are suggested: a center heatmap head to predict center points heatmaps instead of bounding boxes, the center embedding head to predict the embeddings for the center points and the mask feature head to generate mask features. Then, a mask embedding generator takes the predicted semantic classes from the top-K center points from the center heatmap head and passes through a class embedding lookup table module to produce the class embeddings. In parallel, it also takes the coordinates of the top-k predicted center points of the center heatmap head and the center embeddings from the Center Embedding Head. The output mask embeddings result from adding and passing through two fully connected layers that reduce the collisions of multiple instances' center points. The final binary masks are the product of the mask features resulting from the mask feature head and the mask embeddings. The MaskConver was tested on the *COCO* validation dataset, reaching a panoptic quality of 53.6% with 19.6 FPS. In a mobile application using as the backbone MobileNet-MH [169], it yields a value of 29.7 PQ with 375 FPS.

8.4 Transformer-based methods

Transformer methods use transformer-based architectures to perform panoptic segmentation. Employing a transformer-based architecture can lead to highly accurate

models by capturing inter-class relationships and global context while also focusing on fine features. Regardless, these mechanisms require more memory, are more computationally demanding, and take longer to train in order to achieve success.

In computer vision, object detection holds significant importance. Progress in this field has also been applied to segmentation. An example is the DETection TRansformer (DETR) model as described by [170]. This model is designed to tackle object detection as a direct set prediction issue. It leverages a CNN to extract features, which are then combined with positional encoding and inputted into a transformer encoder-decoder architecture. While this architecture follows the standard operations of similar pipelines, it stands out due to the transformer decoder’s ability to decode the N object queries in parallel at each decoder layer. The final prediction is generated by a shared feedforward network (FFN) using the output embedding, predicting either detection or the ‘no box’ class. Additionally, Carion et al. incorporated a mask head on top of the decoder outputs, enabling the model to perform panoptic segmentation. When assessed on the *COCO* validation set using ResNet-101, the model achieved a PQ of 45.1%.

[157] created a transformers-based framework to perform panoptic segmentation. The pipeline uses a mask decoder to generate high-quality masks, resorting to multi-scale attention maps. Furthermore, a query decoupling with bipartite matching was designed to treat both the thing and stuff process separately and avoid the inference between them that results in some poor segmentation masks. The post-processing step utilizes both classification probability and predicted output qualities to perform inference with mask-wise merging. The Panoptic SegFormer was challenged on *COCO* and *ADE20K* datasets. On the first test set, Panoptic SegFormer using Swin-L as backbone reached 56.2% PQ, and on the latter test set, it reached 36.4% PQ.

In the transformer framework for panoptic segmentation developed by [115], clustering serves as the foundation. Object queries are established as the center of the clusters, and pixels are then assigned to these clusters based on feature affinity. Finally, cluster centers are included in the cross-attention module. This process results in a more comprehensive feature map that improves the accuracy of predictions. When tested on *COCO* test set, CMT-DeepLab (iter 200k) using Axial-R104-RFN as backbone reached 55.7% PQ, 83.8% SQ and 65.9% RQ.

[171] have improved the MaskFormer architecture by creating a framework that optimizes segmentation tasks, leading to better results and increased computational efficiency. By switching the Transformer module in the decoder from cross-attention to masked attention, the focus turned to the features close to the predicted regions. The Transformer decoder’s mask attention implementation enabled the query features to become trainable while effectively eliminating the need for dropout through strategic reorganization. Additionally, the communication between the pixel decoder and transformer decoder was designed to improve small object prediction masks. This is achieved by feeding the high-resolution feature from a layer of the pixel decoder to a layer of the Transformer decoder. The Mask2Former was evaluated on the *COCO* dataset, where using Swin-L as the backbone, reached 58.3% PQ, 84.1% SQ, and 68.6% RQ.

[104] realized that most of the models using Transformers to do vision tasks were only based on natural language learning (NLP) models and failed to implement a solution that tackles the learning obstacle of images instead of words. Yu et al. created the k-means Mask XFormer to address a problem. They utilized k-means clustering and modified the cross-attention by replacing softmax with argmax. This redesigns the connection between pixel features and object queries. The overall framework comprises a pixel encoder, which can be any backbone architecture, an improved pixel decoder, and the kMax decoders series that create mask embedding vectors and mask prediction classes. The pixel decoder contains a transformer encoder for bettering pixel features and upsampling layers to generate higher-quality features. The kMaX-DeepLab was the product of integrating this framework with DeepLab. kMAX-DeepLab was evaluated on the *COCO* and *Cityscapes* datasets using ResNet-50 and, ConvNeXt as a backbone. While on the *COCO* test set, kMAX-DeepLab had a better PQ result, 58.5% with the ConvNeXt as the backbone, it was with ResNet-50 that it achieved the better trade-off between panoptic quality, 53.4%, and real-time application, 22.8 FPS. As for the *Cityscapes* test set, using ConvNeXt [64] as a backbone, the kMax-DeepLab yields a panoptic quality percentage of 66.2.

[172] pike upon the work made in kMax-DeepLab [104] to create ReMaX, a model that employs training-time relaxation to the mask transformers. They use a semantic branch during training to predict semantic segmentation masks that are then used to calibrate the panoptic prediction, reducing the chance of false positive predictions. This mechanism is called ReMask or Relaxation on Masks. On the other hand, the Relaxation on Classes (ReClass) technique is used on the masks by replacing the one-hot class label with a softened label, which enables the ground-truth labels to have several classes. The ReMaX pipeline was tested on the *COCO* and *Cityscapes* validation dataset. The panoptic quality scores achieved by the ReMax model using the ResNet-50 and MNV3-S backbones are 54.2% at 16.3 FPS and 40.4% at 108.7 FPS, respectively. On the Cityscapes dataset, the ReMax method with the ResNet-50 backbone reached a PQ of 65.4% at 9.0 FPS, while the MNV3-S backbone attained a PQ of 57.7% at 25.6 FPS.

Mask DINO [159] was developed to integrate the tasks of image detection and segmentation. To achieve this, Li et al. built upon the DINO architecture developed by [173] and added a mask prediction branch. They retained the encoder-decoder structure of DINO and utilized multi-scale features and deformable attention to effectively handle objects of different scales. The prediction branch was integrated into the Transformer decoder to carry out mask classification. As the DINO framework formulates each positional query as a 4D anchor box, it cannot achieve pixel-level alignment. To address this limitation, the authors used the Transformer encoder features and backbone to create a pixel embedding map to obtain the mask. Mask DINO incorporates denoising training from DINO and introduces contrast denoising, which uses both positive and negative samples to help the model distinguish between accurate and noisy data. This approach accelerates convergence and improves stability. Additionally, Mask DINO employs a mixed query selection method to initialize positional queries, enhancing the initial positioning of queries for both detection and segmentation tasks. The authors also utilized the look-forward-twice technique, which

iteratively updates and refines the predictions, improving the accuracy of both bounding boxes and segmentation masks. Mask DINO was evaluated on three segmentation tasks. For panoptic segmentation with the Swin-L backbone, it achieved 59.4% PQ and 54.5% AP on the *COCO* validation set. In semantic segmentation on the *ADE20K* validation set using Swin-L as the backbone, it yielded a mIoU of 59.5%.

Providing a new look at the challenge of panoptic segmentation, [174] proposes an architecture that treats panoptic segmentation in a discretized manner. The Pix2Seq-D is a transformer-based encoder-decoder method which performs segmentation at a grid cell level rather than a pixel level, resulting in a discretization of the panoptic task. The model output space is divided into a grid of cells, each representing a discrete spatial location in the image. For each cell, the model predicts the semantic category, and in the case of an object in the cell, it also predicts the instance ID. A diffusion-based approach is employed in the encoder-decoder architecture. The decoder uses a shared backbone for the semantic and instance head followed by transformer encoder layers, resulting in a feature map. The decoder uses TransUNet [175] architecture, which combines U-Net architecture with transformer decoder layers and takes as input the channel-wise concatenation of the feature map and a randomly initialized noisy mask to output the mask prediction. The decoder goes through a process of iterative denoising the analog bits in the noisy panoptic masks, leveraging the knowledge from the Bit Diffusion [176] mask-generator algorithms used in training and inference. The Pix2Seq-D pipeline was tested on the *COCO* dataset using the ResNet-50 backbone, resulting in a 50.2% PQ result.

[116] built upon the Mask2Former [171] pipeline to create the MP-Former framework. Their work aimed to address the inaccuracies and inconsistencies in mask predictions between consecutive decoding layers observed in the Mask2Former model. To achieve this, they proposed a novel training approach known as mask-piloted training. The MP-Former introduces a new methodology within the Transformer decoder, dividing it into two sections: a matching part similar to the one in Mask2Former, and a mask-piloted (MP) part. The MP part provides the Transformer with additional queries in the form of ground truth (GT) class embeddings and GT masks as attention masks. Outputs from the MP and matching parts are independently assigned to GT instances, following the loss design used in Mask2Former. Additionally, GT masks are integrated into multiple layers and interpolated at different resolutions for various decoder layers to refine mask predictions. The introduction of different types of noise in GT masks on the decoder further improves the mask refinement process. Another key innovation is the label-guided training applied to the first layer, utilizing class embeddings of GT categories as queries in the MP part to distinguish instances based on their categories. The MP-Former was evaluated across all three segmentation tasks. When using Swin-L as the backbone, the MP-Former achieved 50.8% AP on instance segmentation and 58.1% PQ on the *COCO* validation set. On the *ADE20K* validation set, the model attained a 56.9% mean mIoU.

Table 4 Experimental results of panoptic segmentation methods on the test set of *COCO* (green), *Cityscapes* (orange) and *ADE20K* (red). ... denotes information not provided. * refers to data not reported. ~ stands for results on the validation set. The *R* designates the ResNet architecture, while the *X* symbolises the ResNeXt architecture. The results are presented in percentage. All relevant data is incorporated within the exploratory framework

Type	Name	Backbone	Parameters	PQ	FPS	PQ	FPS	PQ	FPS
Top-Down	UPSNNet [108]	R101-FPN [37, 155] R101 [37]	46.1M	46.6	*	61.8	*	*	*
	SpatialFlow [161]	R101-DCN [36, 37] R101 [37]	...	47.9	*	62.5 ~	*	*	*
	REFINE [163]	X101-DCN [36, 52]	...	51.5	*	*	*	*	*
	Ada-Segment [164]	R101-FPN-DCN [36, 37, 155] R50 [37]	...	48.5	*	*	*	32.9	*
	EfficientPS [112]		40.9M	*	*	67.1	*	*	*
Bottom-up	DeeperLab [165]	Xception-71 [57]	...	34.3	*	56.5 ~	*	*	*
	Panoptic-DeepLab [109]	Xception-71 [57]	46.7M	41.4	*	65.6	*	*	*
	Axial-DeepLab [110]	Axial-R-L, Axial-R-XL [110]	44.9M/174M	44.2	*	66.6	*	*	*
	PanoNet [111]	ICNet [77]	12M	*	*	55.1	20	*	*
Single-Stage	MaX-DeepLab [166]	MaX-L [166]	451M	51.3	*	*	*	*	*
	K-Net [81]	Swin-L [58]	...	55.2	*	*	*	*	*
	MaskFormer~ [114]	Swin-L [58]	212M	52.7	5.2	*	*	35.7	*
	RT-K-Net ~ [167]		...	*	*	60.2	31.3	*	*
	YOSO ~ [158]	R50 [37]	...	48.4/46.4	23.6/45.6	59.7/52.5	11.1/22.6	38.0	35.4
	MaskConver ~ [168]	R50/MobileNet-MH [37]	57M/3.4M	53.6/29.7	19.6/375	*	*	*	*
Transformer	DETR ~ [170]	R101 [37] Swin-L [58]	60M	45.1	*	*	*	*	*
	Panoptic SegFormer [157]	R50 [37]	...	56.2	*	*	*	36.4 ~	*
	CMT-DeepLab [115]	Axial-R104 [110]	270.3M	55.7	*	64.6	*	*	*
	Mask2Former ~ [114]	Swin-L [58]	216M	58.3	*	66.6	*	48.1 ~	*
	kMAX-DeepLab [104]	ConvNeXt-L [64]	232M	58.5	6.7	68.4 ~	3.1	50.9 ~	4.0
	Mask DINO ~ [159]	Swin-L [58]	223M	59.4	*	*	*	*	*
	Pix2Seq-D ~ [174]	R50 [37]	94.5M	50.2	*	64.0	*	*	*
	MP-Former ~ [116]	Swin-L [58] ConvNeXt-S/MNV3-S [55, 64]	216M	58.1	4.0	67.5	*	49.4	*
	ReMaX ~ [172]	R50/MNV3-S [37, 55] R50 [37]	83M/18.6M	56.6/40.4	16.5/108.7	65.4/57.5	9.0/25.6	43.4	14.4
Instance	SSAP [129]	R101 [37]	...	36.9	*	61.1 ~	*	*	*
	SOLOv2 ~ [145]	R50-FPN [37, 155]	...	42.1	*	*	*	*	*

9 Exploration Framework

A literature review is a valuable tool for researchers to gain a deeper understanding of the subject matter and to identify commonly used models. Many excellent surveys are available that serve as essential references for researchers. However, a gap has been observed in these papers: the lack of visualization and dynamics to help researchers grasp interconnections and explore the models more easily. To address this, a roadmap was set and an exploration framework¹ was designed. The framework is an interactive, filterable mind map that visualizes the segmentation models presented and discussed in this roadmap. This tool allows users to explore the relationships between models, identify the top models for each metric and task, and access the original research with just a click. This tool was designed with both researchers and students in mind, transforming a traditional survey into a dynamic roadmap of the field. The exploration framework developed for this roadmap is implemented as an open-source project and can be accessed on GitHub at <https://github.com/Matilde3Sousa/Roadmap>.

9.1 Development

The tool was initially conceived as a static mind map intended solely for the article, which can be referenced in Appendix A. During its development, several challenges were encountered due to the numerous variables and intricate interconnections among the models; our objective was to create a design that ensured clarity and legibility. Although the end result was satisfactory, it was still lacking in some aspects. The need to create a dynamic and interactive tool was recognized, allowing researchers to better understand the relationships between the model’s performances based on a variety of metrics, specific for each task. This approach would alleviate the need for researchers to sift through tables and surveys, which can be overwhelming for those who are new to the field. Consequently, it was decided to develop a framework to address this need. Searching for existing tools that could address specific requirements revealed that the available options were too general and not specifically designed for image segmentation. None of the tools provided the relationships and functionalities aimed for. Thus, the design of a custom solution began. To represent the models as nodes, established APIs such as *Cytoscape.js* and *WebCola.js* were used.

9.2 Features

The tool consists of several components, including the representation of models, a search bar, buttons, and filters.

For the model representation, the design targeted a layout similar to the mind map developed for this paper. A library called *WebCola.js* was found, which allows objects to be represented as nodes. The models were defined as nodes, and connections were established based on their interrelationships, with arrows indicating the derivation from one model to another. When a user interacts with a node, they can view links to the related GitHub repository and/or journal, along with performance metrics and other pertinent information. The interface allows users to navigate freely among

¹The exploration framework designed for this roadmap is available online at <https://github.com/Matilde3Sousa/Roadmap>.

multiple nodes and their connections. The search bar serves as a tool for users to locate specific models. When the name of a model is typed, the respective node appears, along with any related models. This functionality ensures that users are informed not only about the model of interest but also about any models that are derived from or related to it. Each task and the **Common Deep Networks** have associated buttons. When pressed, these buttons display all models associated with that task. Furthermore, a checkbox appears, allowing users to filter the displayed nodes. If the user ticks this box, two filters will become available, targeting specific metrics related to the task. Consequently, the interface will adjust to showcase only those nodes that have outcomes for that task. Additionally, a third filter will appear, allowing users to filter models by year.

9.2.1 Filters

To create a fair filtering of the models, it was decided to develop a systematic approach for normalization and ranking. It is important to clarify that the methods were not implemented internally; instead, the data provided by the authors in their respective papers served as the basis for the work. The system consists of several stages, with the initial stage focused on normalizing all values using Min-Max scaling. Equation 19, defines it as follows,

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}. \quad (19)$$

where x represents the value of the metric for a specific model. The terms $\min(x)$ and $\max(x)$ refer to the overall minimum and maximum values of that metric across all models evaluated on the dataset, respectively.

This method normalizes all scores into a range of 0 to 1, making it ideal for representation in a slider format, which is intuitive for users. Additionally, this approach preserves the relative difference between models which is fundamental for a fair comparison of performance across different datasets. Some models were found on our paper that only had data on validation sets. Validation scores cannot be directly compared to test scores due to the differing conditions and factors that may cause validation set results to appear better or worse than those on the test set. Furthermore, since the authors did not present results on the test set, it was decided to give a penalty of 5% for these scores. This approach allows to effectively showcase the capabilities of the models while maintaining awareness of their limited comparability. The value of 5% was reached by considering the scoreboards of multiple datasets and reaching a consensus that this conservative adjustment is a fair representation.

After normalizing the scores of each model, it was applied a weighted average based on the dataset coverage to reflect the importance of each dataset. The mathematical definition is presented in equation 20.

$$\text{Final Score}_M = \begin{cases} \frac{\sum_{i=1}^n V_{M,D_i} \times w_{D_i}}{\sum_{i=1}^n w_{D_i}} + \alpha \times (C(M) - 1), & \text{if } C(M) > 1, \\ V_{M,D_j}, & \text{if } C(M) = 1. \end{cases} \quad (20)$$

where M stands for model, D for dataset, V_{M,D_i} the normalized metric score, w_{D_i} the importance (weight) of a specific dataset, $C(M)$ the number of datasets on which the model was evaluated and α the coverage bonus for each additional dataset.

The reason for using this method was that certain models had results across multiple datasets, unlike others that only had single dataset results. Furthermore, it was considered important to incentivize a generalization across datasets; therefore, a 5% coverage bonus was applied for models evaluated on multiple datasets. As a result, models tested on only one dataset are ranked based on their specific performance on that dataset, while those that generalize well across various benchmarks are also recognized, ensuring fairness among all models. The final evaluation consists of two metrics: model speed and performance metrics, including Average Precision (AP) for instance segmentation, mean Intersection over Union (mIoU) for semantic segmentation, and Panoptic Quality (PQ) for panoptic segmentation. This ranking methodology reflects the datasets and models examined in this paper. Additionally, the complexity of the datasets was not explicitly considered. The final filter sorts the models by their release year.

9.3 Use Cases

Example of use cases for the use of the proposed framework.

Users that can be researcher or other, seeking models to test their solutions can quickly engage with the application to find the most appropriate methods in just a few minutes. They can gain an overview of all relevant models for their specific tasks and even discover other models that may not initially intended for that task but that yield good results. Additionally, users can filter the models to achieve the best quality-performance ratio. For those interested in creating their own models, the application allows them to explore the interconnections between models and common deep network architectures they are based on. By applying filters, they can identify which models are performing best across various metrics. A broad perspective on the current state of the field can help researchers identify the most effective mechanisms for developing better models, as well as create new combinations of models or backbone architectures. The visual overview can also help identify research gaps, thereby driving the field forward. For who is entering the field, the tool serves as a resource to understand the relationships and history between models, as well as find the best methods for each task. They can further explore the underlying methodologies by accessing relevant research papers. Additionally, with just a click of a button, they can be redirected to the project's GitHub page (when provided by the authors), providing direct access to the source code and detailed methodological information. Lastly,

since this is not a static tool, it can be continuously updated by the community and improved based on researchers' needs.

9.3.1 Case One: A new model

It is aimed to develop a new model for semantic segmentation. Their first step is to understand the state-of-the-art, including existing solutions, its mechanisms, and backbones. To achieve this, they access our framework and select the semantic segmentation button, which displays models that perform this task. To focus on the most recent advancements, a filter must be applied based on publication year. The tool then presents the nodes of models that match the chosen year. The multiple nodes can be explored by clicking on them and studying the FPS and mIoU metrics. Intrigued by the evolution of *BiSeNetV3* [85], they use the search tool to find the model's relationships with other models and backbones, as illustrated in Figure 12. This search reveals the model's development process, the prior models on which it was based, and its backbone. By repeating this process for several models, the necessary knowledge can be gathered, equipping them with insights into the most used design patterns and newest mechanisms. This informed perspective enables them to combine established techniques with innovative ideas to push semantic segmentation boundaries further.

9.3.2 Case Two: Plant Phenology Evaluation

A research project wants to analyze the phenology of a plant species to understand the impact of climate change. To achieve this, the project uses computer vision techniques to accurately segment distinct plant traits throughout its life cycle. Accurate segmentation is essential, particularly given challenges such as background noise from other similar plants, variable lighting conditions, and the need to detect disease or infestation in specific plant areas. To identify the most suitable segmentation model, the

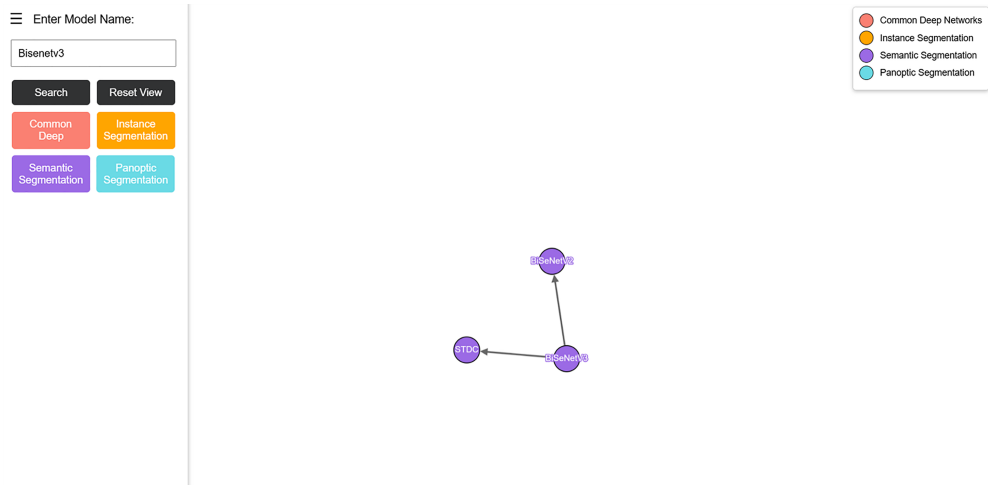


Fig. 12 Illustration of the process for searching a model in our framework, showing the relationships between the searched model and other models

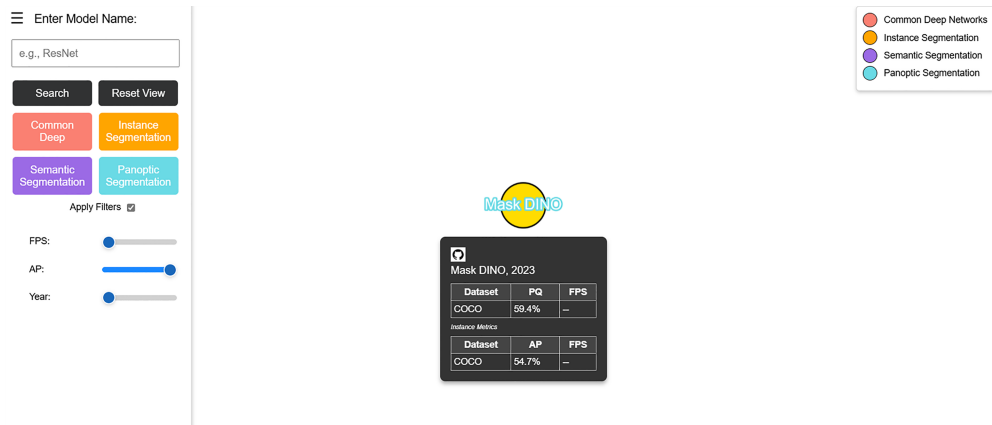


Fig. 13 Representation of the filtering feature and the information retrievable when clicking on the models' node

framework can help. Recognizing that the project requires instance segmentation—for tasks like counting fruits or evaluating flower numbers—the researcher selects the corresponding task within the framework. The tool presents a graph of the nodes for the specific task. Next, the researcher applies a filter (based on AP metric) to pinpoint the models with the highest performance. Given the project's great computational resources, there are no constraints related to processing power. By adjusting the filter, the researcher identifies *Mask DINO* [159] as the model with the best AP. Upon clicking on the models' node, details about other metrics and tasks are displayed, as shown in Figure 13. The model's GitHub repository can be used for an in-depth study of the method and potential use of the code.

10 Challenges and future perspectives

Undoubtedly, the field of image segmentation is continuously advancing and is one of the most significant areas in computer vision when applied to practical applications. To be ready for what lies ahead, it is critical to recognize the potential for development and obstacles that may arise.

- *Computational cost and memory efficiency:* It is evident that newer models are becoming more accurate, especially with the implementation of new transformer-based frameworks. However, this greater accuracy often comes at a higher computational cost. Fortunately, advances in technology have yielded better processors with greater processing capabilities and memory. With these improved accuracy rates, it can be anticipated that segmentation models will play a more significant role in real-world tasks that assist humans. In the years to come, computer vision is expected to contribute to fields such as medical imaging [177, 178] and urban planning [179, 180]. To integrate this technology into our daily lives, including on mobile devices or drones, it is needed less complex architectures that can support the same accuracy rate. Alternatively, techniques like transfer learning or feature

pyramid networks can be employed or use design methods to optimize models, such as compression techniques or network pruning, to achieve a better accuracy versus performance rate.

- *Real-time inference:* In the modern world, achieving success with models requires more than just accuracy. With the rise of autonomous driving, real-time inference has become increasingly critical. The challenge lies in designing models that are both highly accurate and capable of maintaining a reasonable frame rate per second. While many models fall short of achieving real-time inference, there is a growing need to continue striving towards this goal. To tackle this issue, several methods are being employed, including the creation of more lightweight backbones and the use of optimization techniques like atrous-based methods.
- *Datasets:* As more advanced and intricate models are developed, larger and more challenging datasets are required. Networks rely on vast amounts of data to improve their performance, which is why it is crucial to use more complex datasets that include objects with occlusion, objects of varying sizes and formats, difficult scenes where objects blend into the background, and 3D imagery. While some of this complex data already exists, it is often limited in quantity, making it important to increase the size of these datasets with additional images. Some models already use data augmentation methods such as spatial and geometrical transformations. Moreover, researchers are exploring new methods like one-shot learning, zero-data learning and open-vocabulary learning [181–183] to develop models that can surpass the need for massive amounts of data.
- *Unlabelled data:* One challenge that arises is the abundance of unlabelled data. Not only does a vast amount of data need to be collected, but it also needs to be properly labelled. In the past, humans have been responsible for this task, but with the increase in data, there is a shortage of people available to label it. To tackle this issue, researchers are exploring weakly and unsupervised methods. These methods require minimal supervision to learn how to segment objects and do not necessitate pixel-level labels for training. Moreover, they can adapt to situations where annotations may change over time, such as in video or real-time applications. Another viable solution is transfer learning, where a pre-trained model is applied to a task with limited labelled data.

11 Conclusion

This roadmap was conducted regarding the practices of Image Segmentation. Initially, the theoretical aspect of this task and its different branches, including the recent one-shot, single-shot and open vocabulary tasks, has been introduced. Additionally, the types of annotations utilized and their respective applications have been discussed. Furthermore, 17 datasets used in multiple tasks have been presented, including dataset type, scene type, number of images, and annotation type. The roadmap has also explored the most commonly used metrics in image segmentation. In addition, a selection of deep networks that serve as the backbone of multiple image segmentation models has been briefly presented. The main section of the roadmap has reviewed a total of 111 methods, including instance, semantic, and panoptic segmentation tasks.

These models have been organized into subsections based on their taxonomy for better understanding, and their performance in multiple metrics has been compared using various datasets. Furthermore, the development of an exploratory framework was presented. Ultimately, the challenges and future prospects of Image Segmentation have been discussed.

This roadmap confirms that Image Segmentation is still one of the most active branches in computer vision and that its learning still remains fundamental for researchers who aim to address global challenges. It is continuously evolving, with new innovative methods and frameworks emerging every year. Deep learning has a central role in the current paradigm, particularly with the introduction of FCNs and CNNs, which have greatly improved the feasibility of this task. More recently, attention mechanisms and the adaptation of transformers from Large Language Models have further improved the capabilities of models to perform image segmentation. These advancements, coupled with improvements in hardware, have made it possible to apply image segmentation to a wide range of real-life tasks and real-time situations.

When aiming for better results, high-performing models such as transformers and advanced CNNs provide superior accuracy and robustness. These models are suitable for applications where computational resources are abundant, and processing time is not a critical constraint, such as medical image analysis.

In contrast, real-time results demand models optimized for speed and efficiency. These models, often lighter versions of more complex architectures, are designed to operate within the constraints of limited computational resources, making them ideal for applications like real-time video processing in autonomous vehicles, and on-device AI for mobile applications.

Moreover, new solutions for issues like the need for large data are becoming very successful. Research on one-shot, zero-shot, and open vocabulary methods has been increasing exponentially, and it is conceivable that these methods could become the standard in a few years. On the other hand, there is a growing focus on techniques in the domain of unsupervised and weakly supervised learning, which can redefine the future of computer vision. The convergence of one-shot, zero-shot, and open vocabulary methods with unsupervised and weakly supervised learning paradigms is driving advancements in image segmentation. These methods collectively address the challenge of limited annotated data, making it feasible to develop scalable and adaptive models suitable for diverse and dynamic real-world applications.

The final intent of this roadmap is to provide researchers with a deep understanding of image segmentation while also offering practical tools to select the most appropriate model for their specific objectives. The toolkit presented in this work not only includes a comprehensive overview of definitions, building blocks, metrics, and models, but also features an interactive framework that allows users to dynamically explore the interconnections between different approaches and their performance. This framework enables researchers to identify the best options based on the unique conditions of their problems. Ultimately, our innovative tool is designed to advance research in this domain and equip researchers with both foundational knowledge and practical guidance.

Declarations

Funding. This study is within the activities of project Montanha Viva – An intelligent prediction system for decision support in sustainability, project PD21-00009, promoted by PROMOVE program, funded by Fundação La Caixa and supported by Fundação para a Ciência e a Tecnologia and BPI. This research was partially funded by the Fundação para a Ciência e Tecnologia (FCT) and C-MAST (Centre for Mechanical and Aerospace Science and Technologies), under the project UIDB/00151/2020 (<https://doi.org/10.54499/UIDB/00151/2020>; <https://doi.org/10.54499/UIDP/00151/2020>).

Conflict of interest. The authors have no relevant competing interests to declare in relation to the content of this article.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Data Availability. All the datasets used in the paper are publicly available.

Materials availability. Not applicable.

Code availability. All custom scripts and the exploration framework developed are available for public access at: <https://github.com/Matilde3Sousa/Roadmap>.

Author Contributions. The article idea was provided by N. Pereira. M. Sousa and N. Pereira conducted the literature search and data analysis. The article draft was written by M. Sousa, and the work was critically reviewed by N. Pereira and P. Gaspar. All authors have read and approved the final manuscript.

Appendix A Mind Map

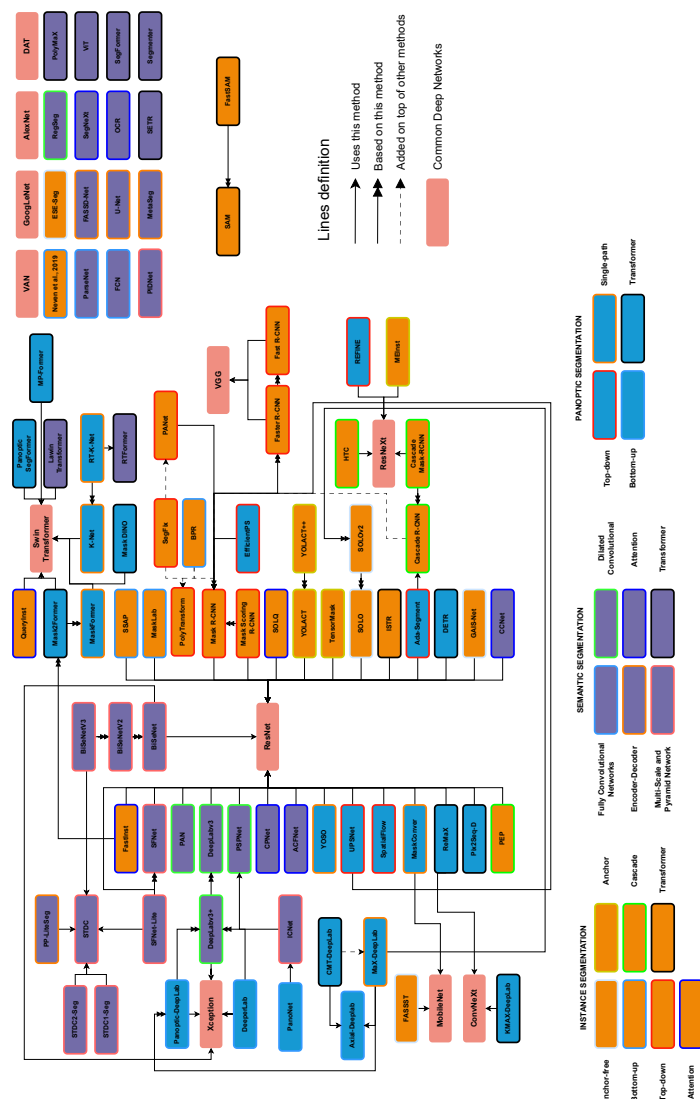


Fig. 14 A mind map illustrating the interconnections and comprehensive overview of methods explored in this roadmap’s review. The approaches without any depicted connections are utilizing an architecture that was not analyzed in this roadmap’s review, or are serving as their own backbone.

References

- [1] Papert, S.: The summer vision project. (1966). <https://api.semanticscholar.org/CorpusID:60684578>
- [2] Szeliski, R.: Computer Vision: Algorithms and Applications. Springer, ??? (2022). <https://doi.org/10.1007/978-3-030-34372-9>
- [3] Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1209–1218 (2018)
- [4] Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H.: Focalclick: Towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1300–1309 (2022)
- [5] Miller, E.G.: Learning from one example in machine vision by sharing probability densities. PhD thesis, Massachusetts Institute of Technology (2002)
- [6] Larochelle, H., Erhan, D., Bengio, Y.: Zero-data learning of new tasks. In: Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2. AAAI’08, pp. 646–651. AAAI Press, ??? (2008)
- [7] Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7086–7096 (2022)
- [8] Wu, J., Li, X., Xu, S., Yuan, H., Ding, H., Yang, Y., Li, X., Zhang, J., Tong, Y., Jiang, X., Ghanem, B., Tao, D.: Towards open vocabulary learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(7), 5092–5113 (2024) <https://doi.org/10.1109/TPAMI.2024.3361862>
- [9] Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A.: Open vocabulary scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
- [10] Ghiasi, G., Gu, X., Cui, Y., Lin, T.-Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European Conference on Computer Vision, pp. 540–557 (2022). Springer
- [11] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3213–3223 (2016)
- [12] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Fleet, D.,

- Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014, pp. 740–755. Springer, Cham (2014)
- [13] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* **88**(2), 303–338 (2010) <https://doi.org/10.1007/s11263-009-0275-4>
 - [14] Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* **32**(11), 1231–1237 (2013) <https://doi.org/10.1177/0278364913491297>
 - [15] Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* **30**(2), 88–97 (2009) <https://doi.org/10.1016/j.patrec.2008.04.005> . Video-based Object and Event Analysis
 - [16] Brostow, G.J., Shotton, J., Fauqueur, J., Cipolla, R.: Segmentation and recognition using structure from motion point clouds. In: *ECCV* (1), pp. 44–57 (2008)
 - [17] Neuhold, G., Ollmann, T., Rota Bulò, S., Kotschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4990–4999 (2017)
 - [18] Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) *Computer Vision – ECCV 2016*, pp. 102–118. Springer, Cham (2016)
 - [19] Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364 (2019)
 - [20] Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with kins dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023 (2019)
 - [21] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., *et al.*: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision* **128**(7), 1956–1981 (2020) <https://doi.org/10.1007/s11263-020-01316-z>
 - [22] Madsen, S.L., Mathiassen, S.K., Dyrmann, M., Laursen, M.S., Paz, L.C., Jørgensen, R.N.: Open plant phenotype database of common weeds in denmark. *Remote Sensing* **12** (2020) <https://doi.org/10.3390/RS12081246>
 - [23] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing

- through ade20k dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
- [24] Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision* **127**(3), 302–321 (2019) <https://doi.org/10.1007/s11263-018-1140-0>
 - [25] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4015–4026 (2023)
 - [26] Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1743–1751 (2019). IEEE
 - [27] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2636–2645 (2020)
 - [28] Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
 - [29] Shaik, F.A., Reddy, A., Billa, N.R., Chaudhary, K., Manchanda, S., Varma, G.: Idd-aw: a benchmark for safe and robust segmentation of drive scenes in unstructured traffic and adverse weather. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4614–4623 (2024)
 - [30] Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. *International journal of computer vision* **111**, 98–136 (2015) <https://doi.org/10.1007/s11263-014-0733-5>
 - [31] Hurtado, J.V., Valada, A.: Chapter 12 - semantic scene segmentation for robotics. In: Iosifidis, A., Tefas, A. (eds.) *Deep Learning for Robot Perception and Cognition*, pp. 279–311. Academic Press, ??? (2022). <https://doi.org/10.1016/B978-0-32-385787-1.00017-8>. <https://www.sciencedirect.com/science/article/pii/B9780323857871000178>
 - [32] Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: European Conference on Computer Vision, pp. 340–353 (2012). Springer
 - [33] Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation.

- In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9404–9413 (2019)
- [34] Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact++ better real-time instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(2), 1108–1121 (2022) <https://doi.org/10.1109/tpami.2020.3014297>
 - [35] Guo, Y., Li, Y., Wang, L., Rosing, T.: Depthwise convolution is all you need for learning multiple visual domains. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8368–8375 (2019)
 - [36] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773 (2017)
 - [37] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
 - [38] Bahdanau, D., Cho, K., Bengio, Y.: Neural Machine Translation by Jointly Learning to Align and Translate (2016). <https://arxiv.org/abs/1409.0473>
 - [39] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015). PMLR
 - [40] Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (2015). <https://arxiv.org/abs/1409.1556>
 - [41] Luong, M.-T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation (2015). <https://arxiv.org/abs/1508.04025>
 - [42] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
 - [43] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., ??? (2017)
 - [44] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (2021). <https://arxiv.org/abs/2010.11929>

- [45] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., ??? (2012)
- [46] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015)
- [47] Lin, M., Chen, Q., Yan, S.: Network In Network (2014). <https://arxiv.org/abs/1312.4400>
- [48] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826 (2016)
- [49] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31 (2017)
- [50] He, K., Sun, J.: Convolutional neural networks at constrained time cost. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5353–5360 (2015)
- [51] Bodhwani, V., Acharjya, D.P., Bodhwani, U.: Deep residual networks for plant identification. *Procedia Computer Science* **152**, 186–194 (2019) <https://doi.org/10.1016/j.procs.2019.05.042>. International Conference on Pervasive Computing Advances and Applications- PerCAA 2019
- [52] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500 (2017)
- [53] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017)
- [54] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520 (2018)
- [55] Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., *et al.*: Searching for mobilenetv3. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324 (2019)

- [56] Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., Aggarwal, V., Zhu, T., Moro, D., Howard, A.: MobileNetV4 – Universal Models for the Mobile Ecosystem (2024). <https://arxiv.org/abs/2404.10518>
- [57] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
- [58] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021)
- [59] Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., Hu, S.-M.: Visual attention network. Computational Visual Media **9**(4), 733–752 (2023) <https://doi.org/10.1007/s41095-023-0364-2>
- [60] Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4794–4803 (2022)
- [61] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
- [62] Cai, Z., Vasconcelos, N.: Cascade r-cnn: High quality object detection and instance segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(5), 1483–1498 (2021) <https://doi.org/10.1109/TPAMI.2019.2956516>
- [63] Xia, Z., Pan, X., Song, S., Li, L.E., Huang, G.: DAT++: Spatially Dynamic Vision Transformer with Deformable Attention (2023). <https://arxiv.org/abs/2309.01430>
- [64] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986 (2022)
- [65] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2019). <https://arxiv.org/abs/1711.05101>
- [66] Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D.: Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3523–3542 (2022) <https://doi.org/10.1109/TPAMI.2021.3059968>
- [67] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic

- p segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
- [68] Liu, W., Rabinovich, A., Berg, A.C.: ParseNet: Looking Wider to See Better (2015). <https://arxiv.org/abs/1506.04579>
 - [69] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer, Cham (2015)
 - [70] Rosas-Arias, L., Benitez-Garcia, G., Portillo-Portillo, J., Sánchez-Pérez, G., Yanai, K.: Fast and accurate real-time semantic segmentation with dilated asymmetric convolutions. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 2264–2271 (2021). IEEE
 - [71] Peng, J., Liu, Y., Tang, S., Hao, Y., Chu, L., Chen, G., Wu, Z., Chen, Z., Yu, Z., Du, Y., et al.: Pp-liteseg: A superior real-time semantic segmentation model. arXiv preprint arXiv:2204.02681 (2022)
 - [72] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., Yan, S.: Metaformer is actually what you need for vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10819–10829 (2022)
 - [73] Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring r-cnn. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6409–6418 (2019)
 - [74] Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6399–6408 (2019)
 - [75] Kang, B., Moon, S., Cho, Y., Yu, H., Kang, S.-J.: Metaseg: Metaformer-based global contexts-aware network for efficient semantic segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 434–443 (2024)
 - [76] Guo, M.-H., Lu, C.-Z., Hou, Q., Liu, Z., Cheng, M.-M., Hu, S.-m.: Segnext: Rethinking convolutional attention design for semantic segmentation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 1140–1156. Curran Associates, Inc., ??? (2022)
 - [77] Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 405–420 (2018)

- [78] Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S., Tong, Y.: Semantic flow for fast and accurate scene parsing. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *Computer Vision – ECCV 2020*, pp. 775–793. Springer, Cham (2020)
- [79] Li, X., Zhou, Y., Pan, Z., Feng, J.: Partial order pruning: For best speed/accuracy trade-off in neural architecture search. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019)
- [80] Li, X., Zhang, J., Yang, Y., Cheng, G., Yang, K., Tong, Y., Tao, D.: SFNet: Faster, Accurate, and Domain Agnostic Semantic Segmentation via Semantic Flow (2022)
- [81] Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems* (2021). <https://openreview.net/forum?id=uDeDDoFOEpj>
- [82] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 325–341 (2018)
- [83] Yu, C., Gao, C., Wang, J., Yu, G., Shen, C., Sang, N.: Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision* **129**, 3051–3068 (2021) <https://doi.org/10.1007/s11263-021-01515-2>
- [84] Fan, M., Lai, S., Huang, J., Wei, X., Chai, Z., Luo, J., Wei, X.: Rethinking bisenet for real-time semantic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9716–9725 (2021)
- [85] Tsai, T.-H., Tseng, Y.-W.: Bisenet v3: Bilateral segmentation network with coordinate attention for real-time semantic segmentation. *Neurocomputing* **532**, 33–42 (2023) <https://doi.org/10.1016/j.neucom.2023.02.025>
- [86] Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722 (2021)
- [87] Xu, J., Xiong, Z., Bhattacharyya, S.P.: Pidnet: A real-time semantic segmentation network inspired by pid controllers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19529–19539 (2023)
- [88] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890 (2017)

- [89] Chen, L.-C., Papandreou, G., Schroff, F., Adam, H.: Rethinking Atrous Convolution for Semantic Image Segmentation (2017). <https://arxiv.org/abs/1706.05587>
- [90] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)
- [91] Li, H., Xiong, P., An, J., Wang, L.: Pyramid Attention Network for Semantic Segmentation (2018). <https://arxiv.org/abs/1805.10180>
- [92] Gao, R.: Rethinking dilated convolution for real-time semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 4675–4684 (2023)
- [93] Zhang, F., Chen, Y., Li, Z., Hong, Z., Liu, J., Ma, F., Han, J., Ding, E.: Acfnnet: Attentional class feature network for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6798–6807 (2019)
- [94] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 603–612 (2019)
- [95] Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, pp. 173–190 (2020). Springer
- [96] Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W., Xiao, B.: Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**(10), 3349–3364 (2021) <https://doi.org/10.1109/TPAMI.2020.2983686>
- [97] Yu, C., Wang, J., Gao, C., Yu, G., Shen, C., Sang, N.: Context prior for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12416–12425 (2020)
- [98] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., *et al.*: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
- [99] Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for

- semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272 (2021)
- [100] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 12077–12090. Curran Associates, Inc., ??? (2021)
 - [101] Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision Transformer Adapter for Dense Predictions (2023). <https://arxiv.org/abs/2205.08534>
 - [102] Yan, H., Zhang, C., Wu, M.: Lawin Transformer: Improving Semantic Segmentation Transformer with Multi-Scale Representations via Large Window Attention (2023). <https://arxiv.org/abs/2201.01615>
 - [103] Wang, J., Gou, C., Wu, Q., Feng, H., Han, J., Ding, E., Wang, J.: Rtformer: Efficient design for real-time semantic segmentation with transformer. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 7423–7436. Curran Associates, Inc., ??? (2022)
 - [104] Yu, Q., Wang, H., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C.: k-means mask transformer. In: European Conference on Computer Vision, pp. 288–307 (2022). Springer
 - [105] Yang, X., Yuan, L., Wilber, K., Sharma, A., Gu, X., Qiao, S., Debats, S., Wang, H., Adam, H., Sirotenko, M., Chen, L.-C.: Polymax: General dense prediction with mask transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1050–1061 (2024)
 - [106] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: Computer Vision—ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part V 12, pp. 746–760 (2012). Springer
 - [107] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., *et al.*: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19175–19186 (2023)
 - [108] Xiong, Y., Liao, R., Zhao, H., Hu, R., Bai, M., Yumer, E., Urtasun, R.: Upsnet: A unified panoptic segmentation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8818–8826 (2019)

- [109] Cheng, B., Collins, M.D., Zhu, Y., Liu, T., Huang, T.S., Adam, H., Chen, L.-C.: Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In: CVPR (2020)
- [110] Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C.: Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation (2020). <https://arxiv.org/abs/2003.07853>
- [111] Chen, X., Wang, J., Hebert, M.: PanoNet: Real-time Panoptic Segmentation through Position-Sensitive Feature Embedding (2020). <https://arxiv.org/abs/2008.00192>
- [112] Mohan, R., Valada, A.: Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision* **129**(5), 1551–1579 (2021) <https://doi.org/10.1007/s11263-021-01445-z>
- [113] Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 418–434 (2018)
- [114] Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 17864–17875. Curran Associates, Inc., ??? (2021)
- [115] Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C.: Cmt-deeplab: Clustering mask transformers for panoptic segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2560–2570 (2022)
- [116] Zhang, H., Li, F., Xu, H., Huang, S., Liu, S., Ni, L.M., Zhang, L.: Mp-former: Mask-piloted transformer for image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18074–18083 (2023)
- [117] Gu, W., Bai, S., Kong, L.: A review on 2d instance segmentation based on deep neural networks. *Image and Vision Computing* **120**, 104401 (2022) <https://doi.org/10.1016/j.imavis.2022.104401>
- [118] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587 (2014)
- [119] Girshick, R.: Fast r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2015)

- [120] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **39**(6), 1137–1149 (2017) <https://doi.org/10.1109/TPAMI.2016.2577031>
- [121] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8759–8768 (2018)
- [122] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141 (2018)
- [123] Zeng, X., Ouyang, W., Yan, J., Li, H., Xiao, T., Wang, K., Liu, Y., Zhou, Y., Yang, B., Wang, Z., Zhou, H., Wang, X.: Crafting gbd-net for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **40**(9), 2109–2123 (2018) <https://doi.org/10.1109/TPAMI.2017.2745563>
- [124] Liang, J., Homayounfar, N., Ma, W.-C., Xiong, Y., Hu, R., Urtasun, R.: Poly-transform: Deep polygon transformer for instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9131–9140 (2020)
- [125] Yuan, Y., Xie, J., Chen, X., Wang, J.: Segfix: Model-agnostic boundary refinement for segmentation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) *Computer Vision – ECCV 2020*, pp. 489–506. Springer, Cham (2020)
- [126] Chen, L.-C., Hermans, A., Papandreou, G., Schroff, F., Wang, P., Adam, H.: Masklab: Instance segmentation by refining object detection with semantic and direction features. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4013–4022 (2018)
- [127] Neven, D., Brabandere, B.D., Proesmans, M., Gool, L.V.: Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. In: *Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition*, pp. 8837–8845 (2019)
- [128] Romera, E., Álvarez, J.M., Bergasa, L.M., Arroyo, R.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems* **19**(1), 263–272 (2018) <https://doi.org/10.1109/TITS.2017.2750080>
- [129] Gao, N., Shan, Y., Wang, Y., Zhao, X., Yu, Y., Yang, M., Huang, K.: Ssap: Single-shot instance segmentation with affinity pyramid. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 642–651 (2019)
- [130] Tang, C., Chen, H., Li, X., Li, J., Zhang, Z., Hu, X.: Look closer to segment

- better: Boundary patch refinement for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13926–13935 (2021)
- [131] Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3150–3158 (2016)
 - [132] Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)
 - [133] Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., *et al.*: Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4974–4983 (2019)
 - [134] Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J.: Dual path networks. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., ??? (2017)
 - [135] Sun, S., Pang, J., Shi, J., Yi, S., Ouyang, W.: Fishnet: A versatile backbone for image, region, and pixel level prediction. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., ??? (2018)
 - [136] Su, J., Yin, R., Chen, X., Luo, J.: Perceive, excavate and purify: A novel object mining framework for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3580–3589 (2023)
 - [137] Fang, Y., Yang, S., Wang, X., Li, Y., Fang, C., Shan, Y., Feng, B., Liu, W.: Instances as queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6910–6919 (2021)
 - [138] Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., *et al.*: Sparse r-cnn: End-to-end object detection with learnable proposals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14454–14463 (2021)
 - [139] Dong, B., Zeng, F., Wang, T., Zhang, X., Wei, Y.: Solq: Segmenting objects by learning queries. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, pp. 21898–21909. Curran Associates, Inc., ??? (2021)
 - [140] He, J., Li, P., Geng, Y., Xie, X.: Fastinst: A simple query-based model for real-time instance segmentation. In: Proceedings of the IEEE/CVF Conference on

- [141] Chen, X., Girshick, R., He, K., Dollár, P.: Tensormask: A foundation for dense object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2061–2069 (2019)
- [142] Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: Yolact: Real-time instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9157–9166 (2019)
- [143] Zhang, R., Tian, Z., Shen, C., You, M., Yan, Y.: Mask encoding for single shot instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10226–10235 (2020)
- [144] Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L.: Solo: Segmenting objects by locations. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020, pp. 649–665. Springer, Cham (2020)
- [145] Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: Solov2: Dynamic and fast instance segmentation. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 17721–17732. Curran Associates, Inc., ??? (2020)
- [146] Xu, W., Wang, H., Qi, F., Lu, C.: Explicit shape encoding for real-time instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5168–5177 (2019)
- [147] Redmon, J., Farhadi, A.: YOLOv3: An Incremental Improvement (2018). <https://arxiv.org/abs/1804.02767>
- [148] Wu, C.-Y., Hu, X., Happold, M., Xu, Q., Neumann, U.: Geometry-Aware Instance Segmentation with Disparity Maps (2024). <https://arxiv.org/abs/2006.07802>
- [149] Cheng, Y., Lin, R., Zhen, P., Hou, T., Ng, C.W., Chen, H.-B., Yu, H., Wong, N.: Fastsst: Fast attention based single-stage segmentation net for real-time instance segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2210–2218 (2022)
- [150] Hu, J., Cao, L., Lu, Y., Zhang, S., Wang, Y., Li, K., Huang, F., Shao, L., Ji, R.: ISTR: End-to-End Instance Segmentation with Transformers (2021). <https://arxiv.org/abs/2105.00637>
- [151] Hu, J., Lu, Y., Zhang, S., Cao, L.: Istr: Mask-embedding-based instance segmentation transformer. IEEE Transactions on Image Processing **33**, 2895–2907 (2024) <https://doi.org/10.1109/TIP.2024.3385980>

- [152] Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast Segment Anything (2023)
- [153] Jocher, G., Chaurasia, A., Qiu, J.: Ultralytics YOLOv8 (2023). <https://github.com/ultralytics/ultralytics>
- [154] Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: European Conference on Computer Vision, pp. 280–296 (2022). Springer
- [155] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
- [156] He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M.: Bag of tricks for image classification with convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 558–567 (2019)
- [157] Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., Lu, T.: Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1280–1289 (2022)
- [158] Hu, J., Huang, L., Ren, T., Zhang, S., Ji, R., Cao, L.: You only segment once: Towards real-time panoptic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17819–17829 (2023)
- [159] Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.-Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3041–3050 (2023)
- [160] Li, X., Chen, D.: A survey on deep learning-based panoptic segmentation. *Digital Signal Processing* **120**, 103283 (2022) <https://doi.org/10.1016/j.dsp.2021.103283>
- [161] Chen, Q., Cheng, A., He, X., Wang, P., Cheng, J.: Spatialflow: Bridging all tasks for panoptic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* **31**(6), 2288–2300 (2021) <https://doi.org/10.1109/TCSVT.2020.3020257>
- [162] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

- [163] Ren, J., Yu, C., Cai, Z., Zhang, M., Chen, C., Zhao, H., Yi, S., Li, H.: Refine: Prediction fusion network for panoptic segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* **35**(3), 2477–2485 (2021) <https://doi.org/10.1609/aaai.v35i3.16349>
- [164] Zhang, G., Gao, Y., Xu, H., Zhang, H., Li, Z., Liang, X.: Ada-segment: automated multi-loss adaptation for panoptic segmentation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 3333–3341 (2021)
- [165] Yang, T.-J., Collins, M.D., Zhu, Y., Hwang, J.-J., Liu, T., Zhang, X., Sze, V., Papandreou, G., Chen, L.-C.: DeeperLab: Single-Shot Image Parser (2019). <https://arxiv.org/abs/1902.05093>
- [166] Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C.: Max-deeplab: End-to-end panoptic segmentation with mask transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5463–5474 (2021)
- [167] Schön, M., Buchholz, M., Dietmayer, K.: Rt-k-net: Revisiting k-net for real-time panoptic segmentation. In: *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–7 (2023). <https://doi.org/10.1109/IV55152.2023.10186625>
- [168] Rashwan, A., Zhang, J., Taalimi, A., Yang, F., Zhou, X., Yan, C., Chen, L.-C., Li, Y.: Maskconver: Revisiting pure convolution model for panoptic segmentation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 851–861 (2024)
- [169] Chu, G., Arikan, O., Bender, G., Wang, W., Brighton, A., Kindermans, P.-J., Liu, H., Akin, B., Gupta, S., Howard, A.: Discovering multi-hardware mobile models via architecture search. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3022–3031 (2021)
- [170] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *European Conference on Computer Vision*, pp. 213–229 (2020). Springer
- [171] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299 (2022)
- [172] Sun, S., WANG, W., Howard, A., Yu, Q., Torr, P., Chen, L.-C.: Remax: Relaxing for better training on efficient panoptic segmentation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*, vol. 36, pp. 73480–73496. Curran Associates, Inc., ??? (2023)

- [173] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.-Y.: DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection (2022). <https://arxiv.org/abs/2203.03605>
- [174] Chen, T., Li, L., Saxena, S., Hinton, G., Fleet, D.J.: A generalist framework for panoptic segmentation of images and videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 909–919 (2023)
- [175] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation (2021). <https://arxiv.org/abs/2102.04306>
- [176] Chen, T., Zhang, R., Hinton, G.: Analog Bits: Generating Discrete Data using Diffusion Models with Self-Conditioning (2023). <https://arxiv.org/abs/2208.04202>
- [177] Aggarwal, M., Tiwari, A.K., Sarathi, M.P., Bijalwan, A.: An early detection and segmentation of brain tumor using deep neural network. BMC Medical Informatics and Decision Making **23**(1), 78 (2023) <https://doi.org/10.1186/s12911-023-02174-8>
- [178] Wu, J., Zhang, Y., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y.: Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation (2023). <https://arxiv.org/abs/2304.12620>
- [179] Kim, H., Lee, J.H., Lee, S.: A hybrid image segmentation method for accurate measurement of urban environments. Electronics **12**(8) (2023) <https://doi.org/10.3390/electronics12081845>
- [180] Guo, Z., Shengoku, H., Wu, G., Chen, Q., Yuan, W., Shi, X., Shao, X., Xu, Y., Shibasaki, R.: Semantic segmentation for urban planning maps based on u-net. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, pp. 6187–6190 (2018). IEEE
- [181] Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., Xie, W.: Learning open-vocabulary semantic segmentation models from natural language supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2935–2944 (2023)
- [182] Luo, H., Bao, J., Wu, Y., He, X., Li, T.: Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: International Conference on Machine Learning, pp. 23033–23044 (2023). PMLR
- [183] Zou, X., Dou, Z.-Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., *et al.*: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15116–15127 (2023)

Phytochemical Profiling and Bioactivity Evaluation of Wild Plants

Alexandra Coimbra ¹, Eugenia Gallardo ^{1,2,3}, Ângelo Luís ¹, Pedro Dinis Gaspar ⁴, Susana Ferreira ¹ and Ana Paula Duarte ^{1*}

¹ RISE-Health – Department of Medical Sciences, Faculty of Health Sciences, University of Beira Interior, Av. Infante D. Henrique, 6200-506 Covilhã, Portugal;

² Laboratório de Fármaco-Toxicologia, UBIMedical, Universidade da Beira Interior, Estrada Municipal 506, 6200-284 Covilhã, Portugal;

³ Grupo de Investigação Sobre Problemas Relacionados Com Toxicofilias, Centro Académico Clínico das Beiras (CACB), Universidade da Beira Interior, Av. Infante D. Henrique, 6200-506 Covilhã, Portugal;

⁴ C-MAST – Center for Mechanical and Aerospace Science and Technologies, University of Beira Interior, Covilhã, Portugal.

* Correspondence: apduarte@fcsaude.ubi.pt

Abstract: Plants used in folk medicine have been increasingly studied to identify their bioactive properties. Therefore, this study aimed to assess the bioactivity of the hydroethanolic extracts of wild plants that were collected in Gardunha Mountain, Portugal. Seven wild plants were studied, *Cistus salviifolius* aerial parts and stems (CSAP, CSS), *Clino-podium vulgare* (CV), *Coincya monensis* flowers and stems (CMF, CMS), *Glandora prostrata* (GP), *Helichrysum stoechas* (HS), *Rubia peregrina* (RP), and *Umbilicus rupestris* flowers and leaves (URF, URL). The phytochemical composition of the extracts was determined by UHPLC–timsTOF–MS, and the total phenolic and flavonoid contents were quantified by spectrophotometric methods. The antioxidant activity, *in vitro* anti-inflammatory activity and the biocompatibility of the extracts were tested, and the antimicrobial activity was evaluated against Gram-positive and Gram-negative bacteria and yeasts. The extracts were predominantly composed of flavonoids and phenolic acids. The extracts demonstrated moderate or very strong antioxidant activity related to scavenging free radicals. Regarding the antimicrobial activity, the extracts exhibited inhibitory effects, particularly against Gram-positive bacteria and yeasts. The CM, RP, and UR extracts showed low cytotoxicity (viability > 70%) in the highest concentration tested. These findings are encouraging for the potential use of these extracts in the discovery of new bioactive compounds.

Keywords: Plant extracts; Phytochemistry; Anti-inflammatory activity; Antioxidant activity; Antimicrobial activity; Cytotoxicity

Academic Editor: First name Last-name

Received: date

Revised: date

Accepted: date

Published: date

Citation: To be added by editorial staff during production.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Since ancient times, plants have played a central role in traditional medicine and remain a primary source of therapeutic agents, especially in developing countries [1,2]. Due to several factors, including the growing demand for natural products, the rise in antibiotic resistance, and the ongoing search for new drugs, plants have increasingly attracted scientific interest, particularly those established in traditional medicine [3–7]. Among these, several plant families are well documented for their ethnobotanical applications

and bioactive properties. Notable examples include: Boraginaceae (anticancer, anti-inflammatory, antimicrobial, antioxidant, and antidiabetic activities, and neuroprotective effect) [8–11]; Brassicaceae (antioxidant, anti-inflammatory, antimicrobial, antiviral, antidiabetic, anticancer, and organ-protective activities) [12–15]; Cistaceae (antioxidant, antimicrobial, anti-inflammatory, and antidiabetic activity) [16,17]; Crassulaceae (anti-inflammatory, antiviral, antimicrobial, insecticidal, anticancer, and antioxidant activity and has wound healing properties) [18–20]; Lamiaceae (antioxidant, anti-inflammatory, anticancer, antimicrobial, and cardioprotective activities and wound-healing and anti-aging potential) [21–25]; and Rubiaceae (antimicrobial, analgesic, antidiabetic, anti-diarrheal, anti-pyretic, anti-inflammatory, antiviral, antiulcer, and antitumor activities) [26–30]. Different representative wild plants from these families can be found in the Gardunha Mountain, located in the central region of Portugal within the Iberian Massif, approximately 17 km west of Castelo Branco [31,32]. This mountain ridge separates the urban areas of Covilhã and Fundão. It is part of an extensive range of ridges stretching approximately 60 km in length, with a predominant NE–SW orientation [31,33]. The soil texture ranges from granular to silt loam, and the summers are extremely hot, with the region experiencing a water deficit of approximately 65%, where winters are mild and almost without snow cover [31,32]. Therefore, the present study aimed to evaluate the chemical composition of wild plants collected in the Gardunha Mountain and to provide a comprehensive understanding of their bioactive properties, as well as their biocompatibility.

2. Results and Discussion

A broad range of plants traditionally used in medicine have been studied by researchers for their health benefits, including antimicrobial and antioxidant activities [34]. Considering the potential of extracts, this work selected seven plants for study, based on their distribution in the Gardunha Mountain, traditional therapeutic uses, and their association with limited knowledge regarding their bioactivity.

2.1. Phytochemical characterisation

The bioactive properties of the plant extracts are closely associated with their chemical composition [35]. Therefore, we started by determining the composition of the extracts, through a preliminary chemical analysis of their total phenolic and flavonoid content. The chemical composition of the extracts was obtained through the determination of total phenolic content using the Folin–Ciocalteu method, and the flavonoid content using the aluminium chloride method (Table 1).

Table 1. Total phenolic and flavonoid contents in extracts. The results were expressed as mg of gallic acid equivalents per gram of extract (mg GAE/g extract) and mg of quercetin equivalents per gram of extract (mg QE/g extract), respectively, as mean \pm standard deviation.

Extracts	Total phenolic content (mg GAE/g extract)	Flavonoid content (mg QE/g extract)
CMF	5.88 \pm 1.26	15.97 \pm 0.79
CMS	12.88 \pm 0.84	14.11 \pm 0.7
CSAP	66.43 \pm 1.45	31.69 \pm 1.16
CSS	62.10 \pm 2.96	15.16 \pm 0.86
CV	64.77 \pm 4.04	19.07 \pm 0.61
GP	86.88 \pm 3.47	18.50 \pm 0.7
HS	57.88 \pm 3.02	50.62 \pm 1.51
RP	24.21 \pm 1.17	15.53 \pm 0.51

URF	48.43 ± 4.00	5.16 ± 0.68
URL	10.43 ± 1.15	6.57 ± 0.32

The extract of *Glandora prostrata* (GP) exhibited the highest total phenolic content, followed by the extracts of *Clinopodium vulgare* (CV), and the stems and aerial parts of *Cistus salviifolius* (CSS and CSAP). Regarding total flavonoid content, the highest concentration was observed in the *Helichrysum stoechas* (HS) extract, followed by the extract obtained from the aerial parts of *C. salviifolius* (CSAP).

The hydroethanolic extract of *C. vulgare* showed a higher total phenolic content compared to ethanol, ethyl acetate, and acetone extracts studied in other works [36]. Among the analysed extracts by Todorova and collaborators (2016), the total aqueous extract of *C. vulgare* exhibited the highest total phenolic content, followed by the leaf aqueous extract, while the butanol extract showed the lowest total phenolic content [37]. Several studies have demonstrated that hydroethanolic extracts (e.g., 50–70% ethanol in water) exhibit significantly higher total phenolic content compared to extracts obtained with less polar solvents [38–40]. Therefore, the choice of plant part may account for the higher phenolic compound content observed when compared to the other plants.

Data on the extracts' composition showed that *C. salviifolius* extracted with methanol acidified with 0.1 % hydrochloric acid (HCl) was characterized by the presence of phenolic acids and their glycosides [41]. The ethanol and water extracts of the same plant had lower levels of total phenols (ranging from 46 to 50 mg GAE/g dry weight (DW)) and the total flavonoid content was lower than the values obtained in the present study but were expressed as catechin equivalents (mg CE/g DW), which may have an impact on the results obtained [42]. The ethanolic and aqueous extracts studied by Hitl and collaborators (2022) showed higher quantities of total phenols and lower total flavonoids as observed in the current investigation [43].

In line with this, when non-targeted metabolomic profiling of the extracts was analysed, it revealed a chemically diverse composition, predominantly comprising phenolic compounds, flavonoids, and coumarins. Compound annotation was performed using a three-tier approach: spectral library matching (SL), comparison with an in-house analyte list (AL), and elemental composition prediction via SmartFormula (SF), all processed using MetaboScape 7.0.1. Across all plant extracts, a total of 72 features were annotated with the highest level of confidence using the combined SL+AL+SF strategy, supported by accurate *m/z*, retention time, and collisional cross section (CCS) values. An additional 657 features were matched using the AL+SF combination, while 116 features were annotated via SL+SF. Furthermore, 2514 features were assigned based solely on SF, providing molecular formula predictions without spectral confirmation.

Comprehensive phytochemical screening of the studied plant extracts by UHPLC–timsTOF–MS enabled the tentative identification of numerous phenolic compounds, with notable variation across species and plant parts.

In *Cistus salviifolius*, the aerial parts showed a presence of flavonoids and phenolic acids, notably gallic acid, neochlorogenic acid, gallic acid 3-O-gallate, and quercetin glycosides, including rutin and arabinoside forms. The stems also contained these compounds but with a slightly different profile, including kaempferol 3-O-glucosyl-rhamnosyl-galactoside and additional coumarins such as scopoletin and coumarin (Table S1).

The extract of *Clinopodium vulgare* featured abundant caffeic acid derivatives, rosmarinic acid, and flavonoid glycosides such as eriocitrin and luteolin derivatives. Noteworthy are also umbelliferone and esculin, suggesting coumarin-type compounds may be important contributors to its phytochemical fingerprint (Table S2).

Coincya monensis exhibited distinct profiles in its flowers and stems. Floral extracts were enriched in *p*-coumaric acid 4-*O*-glucoside, caffeic acid, apigenin 6,8-di-*C*-glucoside, and quercetin 3-*O*-arabinoside. The stems shared these compounds and additionally contained ellagic acid, dihydroquercetin 3-*O*-rhamnoside, and luteolin derivatives (Table S3).

In *Glandora prostrata*, the phytochemical profile was dominated by neochlorogenic acid, kaempferol 3-*O*-glucosyl-rhamnosyl-galactoside, and several flavones such as genistin and luteolin, as well as rosmarinic acid. Rhoifolin derivatives and umbelliferone were also identified (Table S4).

Helichrysum stoechas showed the presence of hydroxycinnamic acid esters, particularly 3-feruloylquinic acid and 3,4-dicaffeoylquinic acid, as well as chlorogenic and neochlorogenic acids. A range of flavonoids, including myricetin, quercetin, and luteolin glycosides, were also detected, along with scopoletin and usnic acid (Table S5).

The extract of *Rubia peregrina* was marked by a diverse flavonoid and phenolic acid profile, including di-*O*-methylbergenin, 3-feruloylquinic acid, quercetin and kaempferol derivatives, as well as anthraquinone-related structures suggested by the presence of emodin-like features (Table S6).

In *Umbilicus rupestris*, both flowers and leaves shared a core set of compounds such as galocatechin, galocatechin 3-*O*-gallate, quercetin 3-*O*-rutinoside, and rosmarinic acid. The flowers were additionally rich in esculetin, ellagic acid, and glycosylated flavones. The leaves contained sinapaldehyde, umbelliferone, and apigenin-*C*-glycosides, consistent with the plant's use in topical formulations (Table S7).

Phenols and flavonoids are plant-derived compounds synthesized in response to various biotic and abiotic stresses. Their levels can vary depending on edaphoclimatic conditions such as light exposure, altitude, ultraviolet radiation, humidity, and temperature [44]. Thus, the differences observed across the various studies may be attributed to these factors, among others.

2.2. Anti-inflammatory activity

Inflammation is a response triggered by infectious agents (like bacteria, viruses, or fungi) or by non-infectious conditions such as tissue injury, cell death, cancer, ischemia, and degeneration, and the use of plants and their derivatives as anti-inflammatory agents dates back to ancient times [45].

Regarding the biological properties of the extracts, their anti-inflammatory activity was evaluated by assessing their ability to inhibit protein denaturation. Although this method is not a direct assay, it is commonly used to estimate the anti-inflammatory potential of plant samples [46].

Based on the results presented in Table 2, the samples with the lowest IC₅₀ values—and thus the highest anti-inflammatory activity—were the *Umbilicus rupestris* leaves extract (URL), followed by the extracts of *H. stoechas* (HR) and *U. rupestris* flowers (URF).

Table 2: Results of anti-inflammatory and antioxidant activity of extracts and standards (results expressed as mean ± standard deviation).

Samples	Anti-inflammatory activity (µg/mL)	Antioxidant activity			
		DPPH method			β-carotene bleaching assay
		IC ₅₀ (µg/mL)	AAI	Antioxidant activity classification	IC ₅₀ (µg/mL)
CMF	371.93 ± 1.55	279.74 ± 57.57	0.21 ± 0.04	Poor	721.55 ± 9.03
CMS	402.71 ± 6.04	307.49 ± 57.57	0.20 ± 0.09	Poor	608.89 ± 9.39

CSAP	745.34 ± 31.96	19.18 ± 5.43	2.93 ± 0.12	Very strong	466.4 ± 10.71
CSS	700.42 ± 14.87	20.72 ± 5.73	2.84 ± 0.28	Very strong	342.82 ± 10.03
CV	309.25 ± 6.85	24.57 ± 7.52	2.32 ± 0.02	Very strong	696.49 ± 25.80
GP	276.31 ± 8.08	25.56 ± 7.80	2.21 ± 0.03	Very strong	352.72 ± 13.39
HS	51.75 ± 3.76	67.83 ± 18.87	0.83 ± 0.04	Moderate	433.09 ± 20.56
RP	304.11 ± 4.91	126.09 ± 78.59	0.55 ± 0.12	Moderate	621.47 ± 20.85
URF	54.79 ± 2.35	69.02 ± 25.72	0.81 ± 0.04	Moderate	335.57 ± 6.76
URL	38.82 ± 2.72	945.01 ± 393.37	0.29 ± 0.11	No activity	505.96 ± 16.77
Acetylsalicylic acid	4.20 ± 1.41	-	-	-	-
Gallic acid	-	3.92 ± 1.26	13.00 ± 0.67	Very strong	-
BHT	-	-	-	-	99.63 ± 10.76

AAI – Antioxidant activity index; BHT – Butylated Hydroxytoluene.

Among the wild mountain plants harvested in the Serra da Gardunha region, *U. rupestris*, particularly its leaves (URL), proved to be the most promising in terms of anti-inflammatory properties. In folk medicine, *U. rupestris* is used against inflammation and irritation and diseases of the skin [47–50]. Regarding the anti-inflammatory activity, Benhouda and Yahia (2015) studied *in vivo* on Wistar rats the anti-inflammatory effects of methanolic extract of *U. rupestris* leaves by using several methods. The methanolic extract induced a significant anti-inflammatory effect after the subcutaneous injection of the carrageenan solution. The effect of this extract on paw edema induced by serotonin and histamine also showed a significant anti-inflammatory activity and dose dependent. This study confirmed the anti-inflammatory properties of *U. rupestris* methanolic extract from the leaves, supporting its traditional medicinal use [50]. These results agree with the ones of the present work, in which the plant *U. rupestris* has anti-inflammatory potential. Also, the anti-inflammatory activity of the *H. stoechas* extract was previously demonstrated, as reported by using different extracts and isolated compounds of *H. stoechas* [51–54]. Indeed, different compounds present in the *U. rupestris* and *H. stoechas* extracts have been shown to have anti-inflammatory activity, such as caffeic acid, coumarin, and rosmarinic acid [55–58]. Although these compounds are present in the other extracts analysed in our study, they did not show notable anti-inflammatory effects. Furthermore, the presence of apigenin and quercetin 3-*O*-rutinoside in the *U. rupestris* extracts, both compounds with previously reported anti-inflammatory activity [59,60], but not in those of *H. stoechas*, may account for the superior anti-inflammatory activity observed in *U. rupestris*, especially in the leaf extract.

2.3. Antioxidant Activity

The evaluation of the antioxidant activity in plant extracts is essential for identifying natural bioactive compounds that may contribute to the prevention of oxidative stress-related diseases and serve as safer alternatives to synthetic antioxidants in pharmaceutical and food applications [61–63].

The antioxidant properties of the extracts were evaluated using two different methodologies. The DPPH (2,2-diphenyl-1-picrylhydrazyl) assay was employed to determine whether the extracts exhibit antioxidant activity with respect to free radical scavenging. Additionally, the β -carotene-bleaching assay was used to assess the ability of the extracts to inhibit lipid peroxidation. The use of multiple methodologies allows for the evaluation

of whether plant extracts possess various mechanisms of action related to antioxidant activity [64].

The DPPH methodology and based on Scherer and Godoy (2009) classification [65], showed that plants exhibiting the best antioxidant activity related to free radical scavenging were *C. salviifolius*, *C. vulgare*, and *G. prostrata*, and in terms of lipid peroxidation inhibition, although all exhibited weak activity, the *U. rupestris* showed a slightly better result (Table 2). It is worth noting that the different parts of these plants also influence their antioxidant activity, with the aerial parts of *C. salviifolius* and the flowers of *U. rupestris* showing the highest antioxidant activity. Petrova et al. (2023) evaluated the total polyphenol and flavonoid contents and antioxidant activity of freeze-dried aqueous extracts from different anatomical parts (leaves, flowers, and stems) of *in vitro* cultivated and wild-growing *C. vulgare* plants, demonstrating how the composition can vary between different parts of the plant and subsequently affect the bioactivity [44]. Several authors showed that different extracts and fractions of *C. salviifolius* may have high antioxidant properties through the radical scavenging activity, with possible correlations between the polyphenol composition and antioxidant activity of the extracts, with the polar extracts being more active than the non-polar ones [41,42,66,67]. In turn, the results reported herein and by other authors, suggest that *C. salviifolius* is a weaker inhibitor of lipid peroxidation [43]. Also, in this work, the scavenging activity of the *C. vulgare* was demonstrated as strong such as previously reported [37,68–71].

The hydroethanolic extracts and decoction of *U. rupestris* demonstrated the ability to scavenge free radicals, inhibit lipid peroxidation, and prevent oxidative damage [72], consistent with the findings of this study on lipid peroxidation inhibition.

Comparing the extracts with the highest antioxidant activity—*C. salviifolius*, *C. vulgare*, and *G. prostrata*—the compounds (+)-galocatechin 3-O-gallate and apigenin-7-glucoside were found in the *C. salviifolius* extracts but not in those of *C. vulgare* and *G. prostrata*. Previous studies have demonstrated the antioxidant activity of these compounds [73,74], suggesting that they may play a role in the slightly better antioxidant potential observed for *C. salviifolius*.

2.4. Antimicrobial activity

Given the global rise of antimicrobial resistance, the investigation of plant extracts as potential sources of novel antimicrobial agents is crucial for the development of safer, more sustainable alternatives or complements to conventional antibiotics. In such a way, plants constitute an underexploited reservoir of antimicrobial agents, with potential applications in medicine, agriculture, and food preservation [75–79].

In the assays conducted to evaluate antimicrobial activity, several Gram-positive and Gram-negative bacterial species, as well as two yeast species, were studied using two different methodologies: disk diffusion susceptibility testing and determination of the minimum inhibitory concentration (MIC) of the extracts.

The results of the disk diffusion assay (Table 3) showed that the extracts of *C. vulgare* (CV), *Coincya monensis* (CMS and CMF), *Rubia peregriana* (RP), and *U. rupestris* (URL and URF) presented no antimicrobial activity against the tested microorganisms using this methodology (data not shown). These findings are broadly consistent with the MIC values obtained (Table 4).

Table 3. Diameters of the inhibition halos (mm) of extracts in bacterial and yeast species are presented as mean \pm standard deviation. Discs with a diameter of 6 mm were used.

Species	Inhibition Zone (10 μ L/Disc)			
	CSAP	CSS	GP	HS
<i>Staphylococcus aureus</i> ATCC 25923	12.67 \pm 0.78	11.72 \pm 0.43	-	24.03 \pm 2.21
<i>Staphylococcus aureus</i> MRSA 05/15	10.05 \pm 0.78	10.77 \pm 0.03	-	22.95 \pm 0.92
<i>Bacillus cereus</i> ATCC 11778	8.52 \pm 0.83	9.99 \pm 0.42	-	21.01 \pm 1.27
<i>Listeria monocytogenes</i> LMG 16779	10.55 \pm 0.33	11.23 \pm 0.57	-	28.91 \pm 1.36
<i>Escherichia coli</i> ATCC 25922	-	-	-	-
<i>Klebsiella pneumoniae</i> ATCC 13883	8.39 \pm 1.24	9.49 \pm 0.43	8.63 \pm 1.32	8.71 \pm 2.18
<i>Pseudomonas aeruginosa</i> ATCC 27853	-	-	-	-
<i>Salmonella</i> Typhimurium ATCC 13311	-	-	-	-
<i>Acinetobacter baumannii</i> AcB 13/10	9.13 \pm 0.11	7.67 \pm 0.73	-	6.52 \pm 0.75
<i>Acinetobacter baumannii</i> LMG 1025	9.11 \pm 0.57	8.73 \pm 0.53	-	6.3 \pm 0.61
<i>Candida albicans</i> ATCC 90028	-	-	-	7.53 \pm 2.64
<i>Candida tropicalis</i> ATCC 750	-	-	-	-

Table 4. Minimum inhibitory concentration (MIC, mg/mL) and minimum lethal concentration (MLC, mg/mL) of extracts on bacterial and yeast species presented as modal values.

Species	MIC (MLC) (mg/mL)									
	CMF	CMS	CSAP	CSS	CV	GP	HS	RP	URF	URL
<i>Staphylococcus aureus</i> ATCC 25923	> 2	> 2	0.5 (2)	0.5 (2)	> 2	2	0.008 (0.03)	> 2	> 2	> 2
<i>Staphylococcus aureus</i> MRSA 05/15	> 2	> 2	0.5 (2)	0.5	> 2	> 2	0.008 (0.03)	> 2	> 2	> 2
<i>Bacillus cereus</i> ATCC 11778	> 2	> 2	0.25	0.5	> 2	> 2	0.008 (0.25)	> 2	> 2	> 2
<i>Listeria monocytogenes</i> LMG 16779	> 2	> 2	0.5 (2)	1	> 2	> 2	0.008 (0.125)	> 2	> 2	> 2
<i>Escherichia coli</i> ATCC 25922	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2
<i>Klebsiella pneumoniae</i> ATCC 13883	> 2	> 2	1	1	2 (2)	1 (1)	2 (2)	> 2	> 2	2
<i>Pseudomonas aeruginosa</i> ATCC 27853	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2
<i>Salmonella</i> Typhimurium ATCC 13311	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2	> 2
<i>Acinetobacter baumannii</i> AcB 13/10	> 2	> 2	2	2 (2)	> 2	> 2	2	> 2	> 2	> 2
<i>Acinetobacter baumannii</i> LMG 1025	> 2	> 2	1	1 (1)	> 2	> 2	2 (2)	> 2	> 2	> 2
<i>Candida albicans</i> ATCC 90028	> 2	> 2	0.03 (2)	0.016	> 2	> 2	0.5 (2)	> 2	0.03	> 2
<i>Candida tropicalis</i> ATCC 750	> 2	> 2	0.25 (0.5)	0.25 (1)	1 (1)	1 (2)	0.5 (1)	> 2	0.25 (1)	> 2

When no MLC value is reported, it indicates that the value is greater than 2 mg/mL.

Among the evaluated extracts, the most promising antimicrobial activity was observed for the *H. stoechas* extract (HS), followed by the extracts of *C. salviifolius* (CSAP and CSS). These extracts exhibited greater activity against Gram-positive bacterial species, as indicated by lower MIC values for these organisms. Regarding the yeast species, the extract of the flowers of *U. rupestris* (URF), along with the extracts of *H. stoechas* (HS) and *C.*

salviifolius (CSAP and CSS), showed higher antimicrobial activity. Concerning the obtained minimum lethal concentrations (MLC) values, most samples exhibited a bacteriostatic effect, with only the CSS, CV, GP, and HS extracts showing a bactericidal activity, primarily against Gram-negative bacteria.

H. stoechas is the plant with the most promising results, exhibiting the largest inhibition zones and the lowest minimum inhibitory concentrations against various Gram-positive and Gram-negative bacteria, as well as yeasts. The obtained results for this extract are in accordance with the results previously reported, demonstrating that different extracts of *H. stoechas* exhibit antimicrobial activity against these microorganisms [80–84].

Regarding *C. salviifolius*, the work of Mastino and collaborators (2021) showed that the butanolic and ethyl acetate fractions of *C. salviifolius* extract demonstrated predominant antimicrobial activity against *S. aureus*, while the aqueous fraction exhibited activity against both *S. aureus* and *Candida* spp.. The authors concluded that these findings indicate that the extraction and partitioning processes influenced both the biological activity and the chemical composition of the *Cistus* extracts [41]. In another study, the ethanolic extract from *Cistus salviifolius* leaves exhibited activity against *L. monocytogenes* [85]. Although the *C. salviifolius* ethanolic extracts studied by Mahmoudi and collaborators, showed higher MIC values, they still demonstrated a bacteriostatic effect [42]. A study on organic extracts of *Cistus salviifolius* identified 2-acetylbenzoic acid as comprising approximately 3.9% of the analysed sample and associated these extracts with antimicrobial activity against ESKAPE clinical isolates [86].

Caffeic acid is a naturally occurring hydroxycinnamic acid, that exhibits a broad spectrum of bioactive properties. This compound has potent antioxidant activity by scavenging free radicals and reducing oxidative stress; can inhibit pro-inflammatory enzymes and cytokines, showing anti-inflammatory activity; possesses antimicrobial activity against various pathogens; and has antiproliferative and pro-apoptotic effects in cancer cell lines, suggesting potential therapeutic applications in oncology [57,58,87,88]. This compound is present in the extracts obtained from the plants *C. salviifolius*, and *H. stoechas* such as also reported in several studies [43,72,83,89,90], suggesting a potential role in the bioactive properties observed. The proposed mechanisms of action for antimicrobial activity involve cell membrane damage, inhibition of DNA or protein synthesis, and induction of oxidative stress [57,88]. Additionally, kaempferol is also present in samples of *C. salviifolius* [91,92]. Kaempferol is a naturally occurring flavonol that exhibits a wide range of pharmacological effects, including antioxidant, anti-inflammatory, antimicrobial, anti-cancer, antidiabetic, neuroprotective, and cardioprotective properties [93–96]. Kaempferol can inhibit the DNA gyrase and DNA helicases in methicillin-resistant *Staphylococcus aureus* [97,98]. In the case of *H. stoechas*, 3,4-dicaffeoylquinic acid, present in its composition, can also be associated with its activity as isomers of caffeoylquinic acids can exhibit antimicrobial and efflux pump inhibitory activity against Gram-positive pathogenic bacteria [99]. Additionally, 4-hydroxyalternariol 9-methyl ether, also present in its composition, has been reported to also possess antibacterial activity [100]. These compounds are not present in the *C. salviifolius* extracts, which may suggest that they contribute to the enhanced antimicrobial activity of *H. stoechas*. Considering these, the antimicrobial activity of the *C. salviifolius*, and *H. stoechas* extracts may possibly be attributed to the presence of these compounds or synergism among them.

2.5. Biocompatibility

Ensuring the safety of plant extracts for human consumption or application is crucial [34]. In such wise, to evaluate the cytotoxic profile of the extracts on human cells, the current investigation was carried out by using cultured Normal Human Dermal Fibroblasts (NHDF) cell line.

The results showed that the extract obtained from the aerial parts of *C. salviifolius* (CSAP) and *H. stoechas* (HS) were the most cytotoxic, reducing cell viability at lower concentrations (Figure 1). NHDF cell viability remained above 75% when exposed to extract concentrations of 1 mg/mL or lower, except for the extracts from the aerial parts of *C. salviifolius* (CSAP) and *H. stoechas* (HS) ($IC_{50} < 1$ mg/mL).

NHDF cells demonstrated high viability (higher than 70 %) when incubated with the extracts of *C. monensis*, *R. peregrina*, and *U. rupestris*. The low cytotoxicity of *U. rupestris* extracts was supported in Human Umbilical Vein Endothelial Cells [101] and in non-tumour primary cell culture from porcine liver [72]. To the best of our knowledge, there are no studies evaluating the cytotoxicity of *C. monensis* and *R. peregrina*.

Despite the cytotoxicity of *C. salviifolius* aerial parts (CSAP) and *H. stoechas* (HS), they also demonstrated antimicrobial activity at concentrations below the IC_{50} found against NHDF cells, supporting their potential application as natural antimicrobial agents.

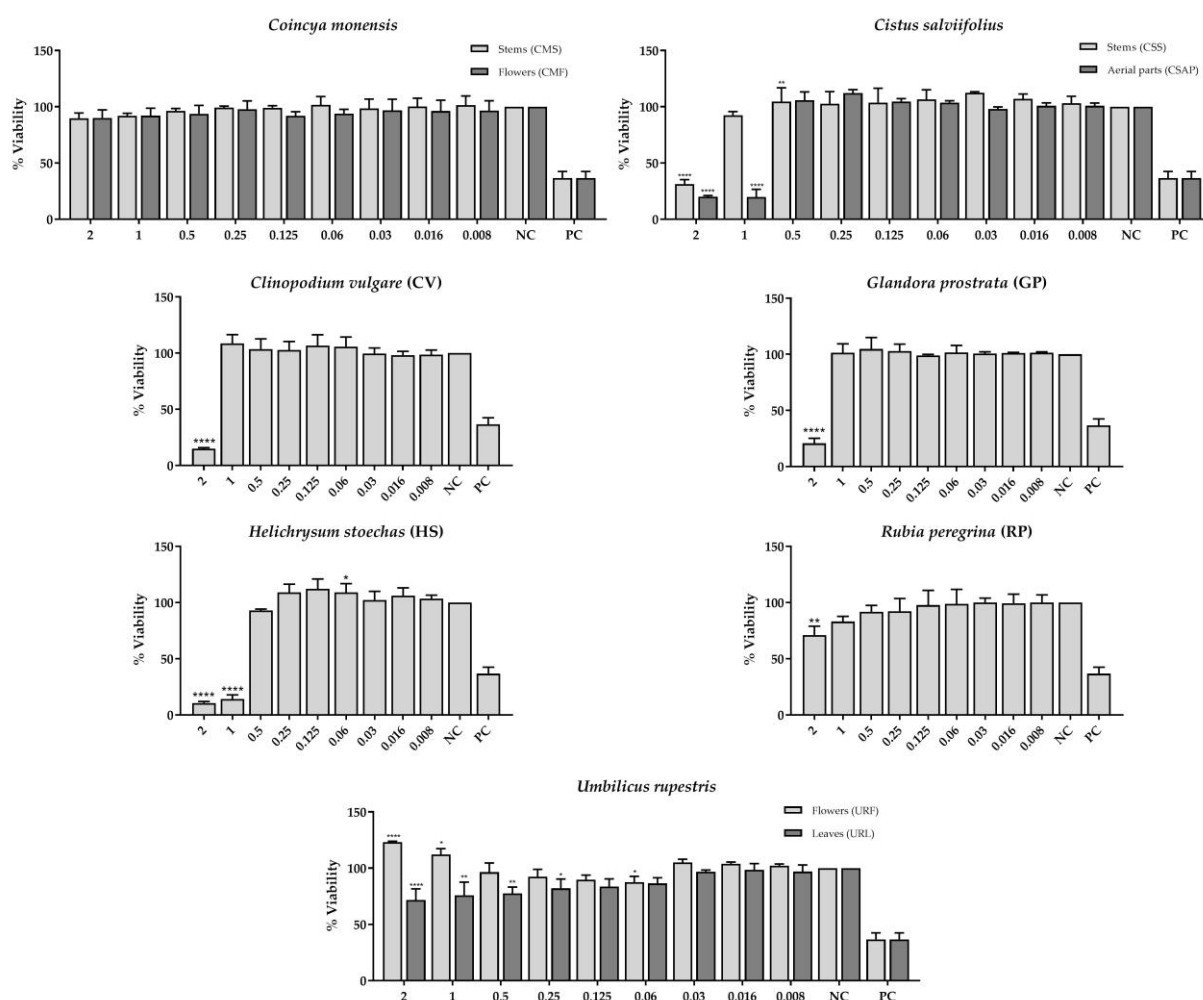


Figure 1. Biocompatibility of extracts for normal human dermal fibroblast (NHDF) cell line measured by MTT assay after 24 h of treatment. Negative control (NC) was performed using untreated cells and cells cultured with Fluorouracil (5-FU, 500 μ g/mL) were used as positive control (PC). Results are expressed as means \pm standard deviation of at least three independent experiments. * ($p < 0.05$); ** ($p < 0.01$); *** ($p < 0.001$); **** ($p < 0.0001$).

3. Materials and Methods

3.1. Collection of plant material

The wild plants *Coinceya monensis* (flowers, CMF or stems, CMS), *Cistus salvifolius* (aerial parts, CSAP or stems, CSS), *Clinopodium vulgare* (aerial parts, CV), *Glandora pros-trata* (aerial parts, GP), *Helichrysum stoechas* (leaves and stems, HS), *Rubia peregrina* (aerial parts, RP), *Umbilicus rupestris* (flowers, URF or leaves, URL), were collected in the north-ern area of Serra da Gardunha, Portugal, during spring of 2023 or spring of 2024. The plant materials were air-dried at room temperature, subsequently milled using a blade disinte-grator and stored under the same conditions until further analysis.

3.2. Extraction

Milled plants (10 g) were extracted with ethanol/water 80:20 (200 mL) in an ultrasonic bath for 1 h with intermittent shaking. At the end of each hour, the extract was removed, and an additional 200 mL of hydroethanolic solution was added. This process was re-peated three times. The extracts were combined and centrifuged at 10000 ×g for 20 min at 4 °C. The supernatant was removed and the ethanolic part evaporated under reduced pressure at 37 °C, and the aqueous part was freezing dried. The extracts were stored at – 20 °C until use.

3.3. Phytochemical analysis

3.3.1. Determination of total phenolic content

The total phenolic content of extracts was determined by the Folin–Ciocalteu colori-metric method accordingly to Luís et al. [102]. Firstly, 450 µL of distilled water was added to 50 µL of each sample or gallic acid solution. Then, 2.5 mL of Folin–Ciocalteu’s reagent (Sigma-Aldrich, USA) (0.2 N) was added and incubated for 5 min before the addition of 2 mL of aqueous Na₂CO₃ (LabKem, USA) (75 g/L). The samples, diluted with methanol (VWR, USA), were then incubated for 90 min at 30 °C. After incubation, the content of total phenolic compounds was determined by colorimetry at 765 nm. A calibration curve was prepared with methanolic solutions of gallic acid (purity 98%, 500–50 mg/L) pur-chased from TCI, Japan ($y = 0.0012x + 0.0257$, $R^2 = 0.99$). The total phenolics were expressed in mg of gallic acid equivalents per gram of extract (mg GAE/g extract) and the analyses were performed in triplicate.

3.3.2. Determination of total flavonoid content

The total flavonoid content of extracts was determined by the aluminium chloride colorimetric method accordingly to Luís et al. [102]. To 500 µL of each extract diluted with methanol, 1.5 mL of methanol, 0.1 mL of aluminium chloride (Scharlau, Spain, 10% w/v), 0.1 mL of 1 M potassium acetate (Fisher Chemical, UK) and 2.8 mL of distilled water were added. Following a 30-minute incubation at room temperature, the absorbance of the so-lutions was measured using a spectrophotometer (UV-6300PC, VWR, USA) at 415 nm. A calibration curve was made with methanolic solutions of quercetin (purity ≥ 95%, 200–12.5 mg/L) purchased from Sigma-Aldrich ($y = 0.0084x + 0.0006$, $R^2 = 0.9993$). The total flavonoids were expressed in mg quercetin equivalents per gram of extract (mg QE/g ex-tract) and the analyses were performed in triplicate.

3.3.3 Analysis of the extracts using Ultra-High Performance Liquid Chromatography cou-pled with trapped ion mobility spectrometry time-of-flight mass spectrometry (UHPLC timsTOF-MS)

To explore the phytochemical composition of the samples, a non-targeted metabolomics strategy was employed using ultra-high-performance liquid chromatography (UHPLC) coupled with trapped ion mobility time-of-flight mass spectrometry (timsTOF-MS; Bruker Daltonics, Germany). The system was equipped with a VIP-HESI electrospray ionisation source to ensure efficient ion generation. Sample extracts were previously reconstituted in the minimum volume of the designated solvent system (ethanol/water 8:2), and 5 µL aliquots were injected into a ZORBAX Eclipse XDB-C18 rapid resolution HD column (2.1 × 100 mm, 1.8 µm; Agilent Technologies, USA).

Chromatographic separation was achieved using a binary solvent system composed of 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B). The gradient programme was as follows: the run began at 2% B (held for 1 minute), increased gradually to 15% B by 7 minutes, and then ramped to 80% B over the next 8 minutes. Subsequently, the gradient rose to 100% B by 20 minutes and was maintained isocratically for 7 minutes. The system then returned to initial conditions, with a re-equilibration period of 2 minutes. The total runtime was 30 minutes, with a constant flow rate of 0.4 mL/min.

Mass spectrometric detection was performed in both positive and negative ionisation modes. Parameters included capillary voltages of ±4500 V and end plate offsets of ±500 V. Nitrogen was used as nebuliser gas (8 bar), drying gas (8 L/min at 240 °C), and sheath gas (4 L/min at 450 °C). Data acquisition ranged from m/z 20–1300, operating in both full-scan MS and MS/MS modes using Parallel Accumulation–Serial Fragmentation (PASEF). Ion mobility data were acquired within a $1/K_0$ range of 0.45–1.45 V·s/cm² with a 100 ms ramp.

Raw data were processed using Bruker's DataAnalysis (v6.1) and MetaboScape (v7.0.1). Feature extraction considered retention time (RT), m/z , and collisional cross section (CCS), with a signal intensity threshold of 10,000 counts. Tentative compound annotation was based on three strategies: spectral library matching (SL), comparison with an in-house analyte list (AL) of known phenolic compounds, and SmartFormula (SF) predictions constrained to CHNOPS atoms.

3.4. Anti-inflammatory activity

The *in vitro* anti-inflammatory activity was performed by evaluating the samples' potential to inhibit protein denaturation according to a previous protocol [103]. Briefly, a solution of bovine serum albumin (BSA, 1% w/v, Sigma-Aldrich, USA) was prepared in phosphate buffer solution (PBS, pH 6.8). The samples were diluted in dimethyl sulfoxide (DMSO) and 1 mL of each were preheated to 37 °C and then, 9 mL of the BSA solution was added. The tubes were subsequently incubated for 10 min at 72 °C, followed by cooling on ice for another 10 minutes. Distilled water was used as a control and acetylsalicylic acid as positive control. Finally, absorbance measurements were performed in triplicate using a spectrophotometer (Helios-Omega, Thermo Scientific, Waltham, MA, USA) at 620 nm. The percentage of inhibition of protein denaturation was quantified using the equation below:

$$\% \text{ Inhibition} = 100 - \left(\left(\text{Abs}_{\text{sample}} \times 100 \right) / \text{Abs}_{\text{control}} \right) \quad (1)$$

where $\text{Abs}_{\text{sample}}$ is the absorbance of each sample and $\text{Abs}_{\text{control}}$ is the absorbance of the control.

3.5. Antioxidant activity

3.5.1. DPPH method

The free radical scavenging activity of the extracts was evaluated using the 2,2-diphenyl-1-picrylhydrazyl radical (DPPH) method as previously reported [102]. The crude

extracts (250–25 mg/L) were compared with gallic acid (10, 25, 50, 75, 100 and 150 mg/L). A calibration curve was constructed with methanolic solutions of DPPH (Sigma-Aldrich, USA, 85–4.25 mg/L) ($y = 0.0084x + 0.0006$, $R^2 = 0.999$). Absorbances were measured at 517 nm and the antioxidant activity was expressed through the Antioxidant Activity Index (AAI), calculated from the following equation:

$$AAI = (\text{final concentration of DPPH in the control sample})/IC_{50} \quad (2)$$

The antioxidant activity of the samples was classified as: poor ($AAI < 0.5$), moderate ($0.5 \leq AAI < 1.0$), strong ($1.0 \leq AAI < 2.0$) or very strong ($AAI \geq 2.0$) accordingly to Scherer and Godoy, [65].

3.5.2. β -carotene-bleaching assay

The ability of extracts to inhibit lipid peroxidation was evaluated by β -carotene bleaching assay [102,104]. For that, methanolic solutions of crude extracts were prepared with concentrations ranging 5–1000 mg/L. Butylated hydroxytoluene (BHT, purity 99%) was purchased from Acros Organics, Belgium, and was used as positive control using the same concentrations. After acquisition of the absorbance values at 470 nm, the percentage of inhibition was calculated:

$$\% \text{ Inhibition} = \left((Abs_{sample}^{t=1h} - Abs_{control}^{t=1h}) / (Abs_{control}^{t=0h} - Abs_{control}^{t=1h}) \right) \times 100 \quad (3)$$

3.6. Antimicrobial activity

3.6.1. Plant extracts, microorganisms and culture media

The extracts were dissolved in dimethyl sulfoxide (DMSO, Fisher Chemical, UK) to a final concentration of 200 mg/mL and stored at -20°C until use.

Gram-positive (*Staphylococcus aureus* ATCC 25923 and MRSA 12/08, *Bacillus cereus* ATCC 11778, and *Listeria monocytogenes* LMG 16779) and Gram-negative bacteria (*Escherichia coli* ATCC 25922, *Klebsiella pneumoniae* ATCC 13883, *Pseudomonas aeruginosa* ATCC 27853, *Salmonella* Typhimurium ATCC 13311, and *Acinetobacter baumannii* LMG 1025 and AcB 13/10), as well as two yeast species (*Candida albicans* ATCC 90028 and *C. tropicalis* ATCC 750) were used.

Müller–Hinton agar (MHA, Biolife, Italy) and Müller–Hinton broth (MHB, Biokar Diagnostics, France) were used for growth and assays with the bacterial species, except for *L. monocytogenes*, for which Tryptone Soy Broth (TSB, VWR, USA) and Tryptone Soy agar (TSA) were used. For yeasts, Sabouraud dextrose agar (SDA, Biokar Diagnostics, France) and Roswell Park Memorial Institute 1640 (RPMI 1640, Biochrom AG, Berlin) supplemented with 3-(N-morpholino)propanesulfonic acid (MOPS, TCI, Japan) were utilized.

All microorganisms used were cryopreserved at -80°C in an appropriate cryoprotective medium containing 20% glycerol (Labchem, Santo Antão do Tojal, Portugal) for long-term storage. For short-term storage during experimental procedures, culture plates were maintained at 4°C . The cultures were subculture onto an appropriate solid medium and incubated at 37°C for 24 h.

3.6.2. Disc-diffusion method

The disc-diffusion method was performed to evaluate the susceptibility of the microorganisms to the extracts as described by Coimbra et al. [105]. Filter paper discs (6 mm) impregnated with 10 μL of each extract (2 mg/disc) were placed on the surface of the inoculated plates. Inhibition zones were measured in millimetres, and results are expressed as mean values \pm standard deviations, based on a minimum of three independent assays.

3.6.3. Determination of the Minimum Inhibitory Concentration (MIC)

The susceptibility of the microorganisms to the extracts was evaluated through the microdilution method accordingly to Coimbra et al. [104]. Briefly, the inoculum prepared by direct suspension in NaCl 0.85% (w/v) was adjusted to 0.5 McFarland and diluted in medium to obtain a final cell concentration of approximately 1×10^6 colony forming unit (CFU) per mL for bacteria and $1\text{--}5 \times 10^3$ CFU/mL for yeasts. The assays were performed with a maximum concentration of 2 mg/mL of the extracts. The 96 well plates were then incubated at 37 °C for 24 h for bacteria and 48 h for yeasts. Then, 30 µL of a 0.01% solution of resazurin (Sigma-Aldrich, USA) were added to the wells, followed by incubation for 2 h (bacteria) or 3 h (yeasts) at 37 °C. To establish minimum bactericidal concentrations (MLC), 10 µL of broth was removed from the well, inoculated on agar plates, and incubated at 37 °C for 24 h for bacteria and 48 h for yeasts. Inhibition of the test microorganism was evidenced by the unchanged blue colour of resazurin. The MLC was defined as the concentration required to kill 99.9% or more of the original microbial population. At least three independent determinations were performed, and the results were presented as modal values.

3.7. Evaluation of extracts biocompatibility

The cytotoxicity of the extracts was evaluated using Normal Human Dermal Fibroblasts (NHDF cells, ATCC, Manassas, VA, USA), initially seeded in 96-well culture plates (100 µL) with 1×10^4 cells/well in Roswell Park Memorial Institute (RPMI-1640, Gibco, USA) supplemented with 10% FBS (PAN Biotech, Germany) and 100 U/mL of penicillin G and 1 mg/mL of streptomycin, and grown with 5% CO₂ at 37 °C for 24 h. Then, the culture medium was removed, and the cells were incubated with several concentrations of extracts for 24 h with RPMI supplemented with 10% FBS without antibiotics. Cells cultured with the well-established chemotherapy agent 5-fluorouracil (5-FU) were used as a positive control, and those with only medium were used as a negative control. The 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT, Sigma-Aldrich, USA) assay was used to evaluate cell metabolic function. For that, the medium was removed, and an RPMI solution with 0.5 mg/mL of MTT (100 µL in each well) was added to each sample. The plates were incubated in the same conditions for 3 h. Then, the pigmented formazan formed was dissolved with 100 µL of DMSO. Afterwards, the absorbance at 570 nm was read in a microplate spectrophotometer, Bio-Rad (Hercules, CA, USA) xMark. The statistical analysis of the results was performed using the one-way ANOVA and Dunnett test using the GraphPad Prism v8.01 software, with a 95% confidence interval, considering the values of $p < 0.05$ as statistically significant.

4. Conclusions

Flavonoids and phenolic acids predominated in the composition of the extracts. Among the wild mountain plants collected in the Serra da Gardunha region, *Umbilicus rupestris*, particularly its leaves, was identified as the most promising plant in terms of anti-inflammatory properties. The plants that exhibited antioxidant activity, regarding free radical scavenging, were *Cistus salvifolius*, and in terms of lipid peroxidation inhibition, *Umbilicus rupestris*. It is worth noting that different plant parts also influenced antioxidant activity, with the aerial parts of *Cistus salvifolius* and the flowers of *Umbilicus rupestris* showing the highest antioxidant potential. Regarding antimicrobial activity, the extracts of *Cistus salvifolius* and *Helichrysum stoechas* showed the most promising results, exhibiting larger inhibition zones and lower minimum inhibitory concentrations against various Gram-positive and Gram-negative bacteria, as well as yeasts. It can be concluded that extracts from these plants demonstrated promising bioactive activities, emphasizing

the *C. salviifolius* and *H. stoechas*. These results are promising, indicating the potential use of these extracts in the discovery of new bioactive compounds, highlighting the knowledge attributed to traditional medicine.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/doi/s1>, **Table S1.** Representative compounds identified in *Cistus salviifolius* extracts using UHPLC–timsTOF–MS, based on combined annotation by spectral library matching, in-house analyte list, and SmartFormula. Identification supported by accurate mass, retention time, and collisional cross section. **Table S2.** Representative compounds identified in *Clino-podium vulgare* extract. **Table S3.** Representative compounds identified in *Coincya monensis* extracts. **Table S4.** Representative compounds identified in *Glandora prostrata* extract. **Table S5.** Representative compounds identified in *Helichrysum stoechas* extract. **Table S6.** Representative compounds identified in *Rubia peregrina* extract. **Table S7.** Representative compounds identified in *Umbilicus rupestris* extracts.

Author Contributions: Conceptualization, Â.L., P.D.G., S.F., and A.P.D; methodology, A.C., Â.L. and E.G.; formal analysis, A.C., and S.F.; writing—original draft preparation, A.C.; writing—review and editing, A.C., Â.L., E.G., P.D.G., S.F., and A.P.D.; supervision, Â.L., S.F. and A.P.D; funding acquisition, Â.L., P.D.G., S.F. and A.P.D.; project administration, P.D.G.. All authors have read and agreed to the published version of the manuscript.

Funding: This research is within the activities of the project Montanha Viva – Sistema Previsional Inteligente de Suporte à Decisão em Sustentabilidade”, project PD21-00009, promoted by PROMOVE program funded by Fundação La Caixa and supported by Fundação para a Ciência e a Tecnologia and BPI. This work was developed within the scope of the CICS-UBI projects UIDB/00709/2020 and UIDP/00709/2020, financed by national funds through the Portuguese Foundation for Science and Technology/MCTES. Pedro Dinis Gaspar acknowledges the support of FCT – Fundação para a Ciência e a Tecnologia, I.P. and Centre for Mechanical and Aerospace Science and Technologies (C-MAST), under the project UIDB/00151/2020 (<https://doi.org/10.54499/UIDB/00151/2020>; <https://doi.org/10.54499/UIDP/00151/2020>).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the text.

Acknowledgments: Alexandra Coimbra is recipient of a research fellowship within the Research project titled “Montanha Viva - Sistema Previsional Inteligente de Suporte à Decisão em Sustentabilidade” (Ref. PD21-00009), funded by Fundação La Caixa.

Conflicts of Interest: The authors declare no conflicts of interest

References

1. Rahayu, Y.Y.S.; Araki, T.; Rosleine, D. Factors Affecting the Use of Herbal Medicines in the Universal Health Coverage System in Indonesia. *J Ethnopharmacol* **2020**, *260*, 112974, doi:10.1016/j.jep.2020.112974.
2. Kim, J.K.; Kim, K.H.; Shin, Y.C.; Jang, B.H.; Ko, S.G. Utilization of Traditional Medicine in Primary Health Care in Low-and Middle-Income Countries: A Systematic Review. *Health Policy Plan* **2020**, *35*, 1070–1083, doi:10.1093/heapol/czaa022.
3. Olas, B.; Róžański, W.; Urbańska, K.; Sławińska, N.; Bryś, M. New Light on Plants and Their Chemical Compounds Used in Polish Folk Medicine to Treat Urinary Diseases. *Pharmaceuticals* **2024**, *17*, 435, doi:10.3390/ph17040435.

4. Zhou, P.; Hu, H.; Wu, X.; Feng, Z.; Li, X.; Tavakoli, S.; Wu, K.; Deng, L.; Luo, H. Botany, Traditional Uses, Phytochemistry, Pharmacological Activities, and Toxicity of the Mangrove Plant *Avicennia Marina*: A Comprehensive Review. *Phytochem Rev* **2025**, doi:10.1007/s11101-025-10080-2. 550–552
5. Vazquez-Marquez, A.M.; Correa-Basurto, J.; Varela-Guerrero, V.; González-Pedroza, M.G.; Zepeda-Gómez, C.; Burrola-Aguilar, C.; Nieto-Trujillo, A.; Estrada-Zúñiga, M.E. Phytoremediation and Ethnomedicinal Potential of *Buddleja*, *Eremophila*, *Scrophularia*, and *Verbascum* Genera Belonging to the Scrophulariaceae Family: A Review. *Futur J Pharm Sci* **2024**, *10*, 173, doi:10.1186/s43094-024-00742-x. 553–556
6. Qin, Y.; Wang, N.; Pan, H.; Lei, X.; Li, X. *Hellenia speciosa*: A Comprehensive Review of Traditional Applications, Phytonutrients, Health Benefits and Safety. *Food Chem* **2025**, *465*, 142003, doi:10.1016/j.foodchem.2024.142003. 557–558
7. Yesilada, E. Scientific Evaluation of the Remedies Used in Turkish Folk Medicine to Treat Possible Viral Infections. *Curr Tradit Med* **2022**, *9*, 1–15, doi:10.2174/2215083809666221227143652. 559–560
8. Chrzanowska, E.; Denisow, B.; Ekiert, H.; Pietrzyk, Ł. Metabolites Obtained from Boraginaceae Plants as Potential Cosmetic Ingredients—A Review. *Molecules* **2024**, *29*, 5088, doi:10.3390/molecules29215088. 561–562
9. Gautam, S.; Lapčík, L.; Lapčíková, B. Pharmacological Significance of Boraginaceae with Special Insights into Shikonin and Its Potential in the Food Industry. *Foods* **2024**, *13*, 1350 doi:10.3390/foods13091350. 563–564
10. Sharma, R.A.; Singh, B.; Singh, D.; Chandrawat, P. Ethnomedicinal, Pharmacological Properties and Chemistry of Some Medicinal Plants of Boraginaceae in India. *J Med Plants Res* **2009**, *3*, 1153–1175. 565–566
11. Zhang, Z.; Bai, J.; Zeng, Y.; Cai, M.; Yao, Y.; Wu, H.; You, L.; Dong, X.; Ni, J. Pharmacology, Toxicity and Pharmacokinetics of Acetylshikonin: A Review. *Pharm Biol* **2020**, *58*, 950–958, doi:10.1080/13880209.2020.1818793. 567–568
12. Cicio, A.; Serio, R.; Zizzo, M.G. Anti-Inflammatory Potential of Brassicaceae-Derived Phytochemicals: *In Vitro* and *In Vivo* Evidence for a Putative Role in the Prevention and Treatment of IBD. *Nutrients* **2023**, *15*, 31, doi:10.3390/nu15010031. 569–570
13. Mattosinhos, P. da S.; Sarandy, M.M.; Novaes, R.D.; Esposito, D.; Gonçalves, R.V. Anti-Inflammatory, Antioxidant, and Skin Regenerative Potential of Secondary Metabolites from Plants of the Brassicaceae Family: A Systematic Review of *In Vitro* and *In Vivo* Preclinical Evidence (Biological Activities Brassicaceae Skin Diseases). *Antioxidants* **2022**, *11*, 1346, doi:10.3390/antiox11071346. 571–574
14. Rahman, M.; Khatun, A.; Liu, L.; Barkla, B.J. Brassicaceae Mustards: Phytochemical Constituents, Pharmacological Effects, and Mechanisms of Action against Human Disease. *Int J Mol Sci* **2024**, *25*, 9039, doi:10.3390/ijms25169039. 575–576
15. Zhang, N.; Jing, P. Anthocyanins in Brassicaceae: Composition, Stability, Bioavailability, and Potential Health Benefits. *Crit Rev Food Sci Nutr* **2022**, *62*, 2205–2220, doi:10.1080/10408398.2020.1852170. 577–578
16. Bedoya, L.M.; Bermejo, P.; Abad, M.J. Anti-Infectious Activity in the Cistaceae Family in the Iberian Peninsula. *Mini Rev Med Chem* **2009**, *9*, 519–525, doi:10.2174/138955709788167600. 579–580
17. Tomou, E.M.; Lytra, K.; Rallis, S.; Tzakos, A.G.; Skaltsa, H. An Updated Review of Genus *Cistus* L. since 2014: Traditional Uses, Phytochemistry, and Pharmacological Properties. *Phytochem Rev* **2022**, *21*, 2049–2087, doi:10.1007/s11101-022-09827-y. 581–582
18. Assis de Andrade, E.; Machinski, I.; Terso Ventura, A.C.; Barr, S.A.; Pereira, A.V.; Beltrame, F.L.; Strangman, W.K.; Williamson, R.T. A Review of the Popular Uses, Anatomical, Chemical, and Biological Aspects of *Kalanchoe* (Crassulaceae): A Genus of Plants Known as “Miracle Leaf.” *Molecules* **2023**, *28*, 5574, doi:10.3390/molecules28145574. 583–585
19. Hassan, M.H.A.; Elwekeel, A.; Moawad, A.; Afifi, N.; Amin, E.; El Amir, D. Phytochemical Constituents and Biological Activity of Selected Genera of Family Crassulaceae: A Review. *S Afr J Bot* **2021**, *141*, 383–404, doi:10.1016/j.sajb.2021.05.016. 586–587
20. Salazar-Gómez, A.; Velo-Silvestre, A.A.; Alonso-Castro, A.J.; Hernández-Zimbrón, L.F. Medicinal Plants Used for Eye Conditions in Mexico—A Review. *Pharmaceuticals* **2023**, *16*, 1432, doi:10.3390/ph16101432. 588–589
21. Diab, F.; Zbeeb, H.; Baldini, F.; Portincasa, P.; Khalil, M.; Vergani, L. The Potential of Lamiaceae Herbs for Mitigation of Overweight, Obesity, and Fatty Liver: Studies and Perspectives. *Molecules* **2022**, *27*, 5043, doi:10.3390/molecules27155043. 590–591
22. Islam, A.K.M.M.; Suttiyut, T.; Anwar, M.P.; Juraimi, A.S.; Kato-Noguchi, H. Allelopathic Properties of Lamiaceae Species: Prospects and Challenges to Use in Agriculture. *Plants* **2022**, *11*, 1478, doi:10.3390/plants11111478. 592–593

23. Kowalczyk, T.; Merecz-Sadowska, A.; Ghorbanpour, M.; Szemraj, J.; Piekarski, J.; Bijak, M.; Śliwiński, T.; Zajdel, R.; Sitarek, P. Enhanced Natural Strength: Lamiaceae Essential Oils and Nanotechnology in *In Vitro* and *In Vivo* Medical Research. *Int J Mol Sci* **2023**, *24*, 15279, doi:10.3390/ijms242015279.
24. Michel, J.; Abd Rani, N.Z.; Husain, K. A Review on the Potential Use of Medicinal Plants From Asteraceae and Lamiaceae Plant Family in Cardiovascular Diseases. *Front Pharmacol* **2020**, *11*, 852, doi:10.3389/fphar.2020.00852.
25. Ramos Da Silva, L.R.; Ferreira, O.O.; Cruz, J.N.; De Jesus Pereira Franco, C.; Oliveira Dos Anjos, T.; Cascaes, M.M.; Almeida Da Costa, W.; Helena De Aguiar Andrade, E.; Santana De Oliveira, M. Lamiaceae Essential Oils, Phytochemical Profile, Antioxidant, and Biological Activities. *Evidence-Based Complementary Altern Med* **2021**, *2021*, 6748052, doi:10.1155/2021/6748052.
26. González-Castelazo, F.; Soria-Jasso, L.E.; Torre-Villalvazo, I.; Cariño-Cortés, R.; Muñoz-Pérez, V.M.; Ortiz, M.I.; Fernández-Martínez, E. Plants of the Rubiaceae Family with Effect on Metabolic Syndrome: Constituents, Pharmacology, and Molecular Targets. *Plants* **2023**, *12*, 3583, doi:10.3390/plants12203583.
27. Jaafar, A.; Zulkiply, M.A.; Mohd Hatta, F.H.; Jahidin, A.H.; Abdul Nasir, N.A.; Hazizul Hasan, M. Therapeutic Potentials of Iridoids Derived from Rubiaceae against *In Vitro* and *In Vivo* Inflammation: A Scoping Review. *Saudi Pharm J* **2024**, *32*, 101876, doi:10.1016/j.jsps.2023.101876.
28. Martins, D.; Nunez, C.V. Secondary Metabolites from Rubiaceae Species. *Molecules* **2015**, *20*, 13422–13495, doi:10.3390/molecules200713422.
29. Roy, D.; Brar, S.; Bhatia, R.; Rangra, N.K. An Insight into the Ethnopharmacological Importance of Indian Subcontinent Medicinal Plant Species of Rubiaceae Family. *Adv Tradit Med* **2023**, *24*, 947–969, doi:10.1007/s13596-023-00714-1.
30. Su, G.Y.; Chen, M.L.; Wang, K.W. Natural New Bioactive Anthraquinones from Rubiaceae. *Mini Rev Org Chem* **2020**, *17*, 872–883, doi:10.2174/1570193X17666200107092510.
31. Pakpahan, O.P.; Moreira, L.; Camelo, A.; Karya, D.; Martins, A.C.; Gaspar, P.D.; Santo, C.E. Evaluation of Comparative Scenarios from Different Sites of Chestnut Production Using Life Cycle Assessment (LCA): Case Study in the Beira Interior Region of Portugal. *Heliyon* **2023**, *9*, e12847, doi:10.1016/j.heliyon.2023.e12847.
32. Carvalho, P.; Nogueira, A.J.A.; Soares, A.M.V.M.; Fonseca, C. Ranging Behaviour of Translocated Roe Deer in a Mediterranean Habitat: Seasonal and Altitudinal Influences on Home Range Size and Patterns of Range Use. *Mammalia* **2008**, *72*, 89–94, doi:10.1515/MAMM.2008.019.
33. Teixeira, J.; Chaminé, H.I.; Carvalho, J.M.; Pérez-Alberti, A.; Rocha, F. Hydrogeomorphological Mapping as a Tool in Groundwater Exploration. *J Maps* **2013**, *9*, 263–273, doi:10.1080/17445647.2013.776506.
34. Silva, B.N.; Cadavez, V.; Ferreira-Santos, P.; Alves, M.J.; Ferreira, I.C.F.R.; Barros, L.; Teixeira, J.A.; Gonzales-Barron, U. Chemical Profile and Bioactivities of Extracts from Edible Plants Readily Available in Portugal. *Foods* **2021**, *10*, 673, doi:10.3390/foods10030673.
35. Sytar, O.; Hemmerich, I.; Zivcak, M.; Rauh, C.; Brestic, M. Comparative Analysis of Bioactive Phenolic Compounds Composition from 26 Medicinal Plants. *Saudi J Biol Sci* **2018**, *25*, 631–641, doi:10.1016/j.sjbs.2016.01.036.
36. Stefanovic, O.; Stankovic, M.S.; Comic, L. *In Vitro* Antibacterial Efficacy of *Clinopodium vulgare* L. Extracts and Their Synergistic Interaction with Antibiotics. *J Med Plants Res* **2011**, *5*, 4074–4079.
37. Todorova, T.; Ventzislavov Bardarov, K.; Miteva, D.; Bardarov, V. DNA-Protective Activities of *Clinopodium vulgare* L. Extracts. *C R Acad Bulg Sci* **2016**, *69*, 1019–1024.
38. Nakilcioglu-Taş, E.; Ötleş, S. Influence of Extraction Solvents on the Polyphenol Contents, Compositions, and Antioxidant Capacities of Fig (*Ficus Carica* L.) Seeds. *An Acad Bras Cienc* **2021**, *93*, e20190526, doi:10.1590/0001-3765202120190526.
39. Xiang, Z.; Liu, L.; Xu, Z.; Kong, Q.; Feng, S.; Chen, T.; Zhou, L.; Yang, H.; Xiao, Y.; Ding, C. Solvent Effects on the Phenolic Compounds and Antioxidant Activity Associated with *Camellia polyodonta* Flower Extracts. *ACS Omega* **2024**, *9*, 27192–27203, doi:10.1021/acsomega.4c01321.

40. Tourabi, M.; Faiz, K.; Ezzouggar, R.; Louasté, B.; Merzouki, M.; Dauelbait, M.; Bourhia, M.; Almaary, K.S.; Siddique, F.; Lyoussi, B.; Derwich, E. Optimization of Extraction Process and Solvent Polarities to Enhance the Recovery of Phytochemical Compounds, Nutritional Content, and Biofunctional Properties of *Mentha longifolia* L. Extracts. *Bioresour Bioprocess* **2025**, *12*, 24, doi:10.1186/s40643-025-00859-8.
41. Mastino, P.M.; Marchetti, M.; Costa, J.; Juliano, C.; Usai, M. Analytical Profiling of Phenolic Compounds in Extracts of Three *Cistus* Species from Sardinia and Their Potential Antimicrobial and Antioxidant Activity. *Chem Biodivers* **2021**, *18*, e2100053, doi:10.1002/cbdv.202100053.
42. Mahmoudi, H.; Aouadhi, C.; Kaddour, R.; Gruber, M.; Zargouni, H.; Zaouali, W.; Ben Hamida, N.; Ben Nasri, M.; Ouerghi, Z.; Hosni, K. Comparison of Antioxidant and Antimicrobial activities of Two Cultivated *Cistus* Species from Tunisia. *Biosci J* **2016**, *32*, 226–237, doi:10.14393/BJ-v32n1a2016-30208.
43. Hitl, M.; Bijelić, K.; Stilinović, N.; Božin, B.; Srđenović-Čonić, B.; Torović, L.; Kladar, N. Phytochemistry and Antihyperglycemic Potential of *Cistus salvifolius* L., Cistaceae. *Molecules* **2022**, *27*, 8003, doi:10.3390/molecules27228003.
44. Petrova, M.; Dimitrova, L.; Dimitrova, M.; Denev, P.; Teneva, D.; Georgieva, A.; Petkova-Kirova, P.; Lazarova, M.; Tasheva, K. Antitumor and Antioxidant Activities of *In Vitro* Cultivated and Wild-Growing *Clinopodium vulgare* L. Plants. *Plants* **2023**, *12*, 1591, doi:10.3390/plants12081591.
45. Azab, A.; Nassar, A.; Azab, A.N. Anti-Inflammatory Activity of Natural Products. *Molecules* **2016**, *21*, 1321, doi:10.3390/molecules21101321.
46. Gunathilake, K.D.P.P.; Ranaweera, K.K.D.S.; Rupasinghe, H.P.V. *In Vitro* Anti-Inflammatory Properties of Selected Green Leafy Vegetables. *Biomedicines* **2018**, *6*, 107, doi:10.3390/biomedicines6040107.
47. Neves, J.M.; Matos, C.; Moutinho, C.; Queiroz, G.; Gomes, L.R. Ethnopharmacological Notes about Ancient Uses of Medicinal Plants in Trás-os-Montes (Northern of Portugal). *J Ethnopharmacol* **2009**, *124*, 270–283, doi:10.1016/j.jep.2009.04.041.
48. Novais, M.H.; Santos, I.; Mendes, S.; Pinto-Gomes, C. Studies on Pharmaceutical Ethnobotany in Arrabida Natural Park (Portugal). *J Ethnopharmacol* **2004**, *93*, 183–195, doi:10.1016/j.jep.2004.02.015.
49. Benhouda, A.; Benhouda, D.; Yahia, M. *In Vivo* Evaluation of Anticryptosporidiosis Activity of the Methanolic Extract of the Plant *Umbilicus rupestris*. *Biodiversitas* **2019**, *20*, 3478–3483, doi:10.13057/biodiv/d201203.
50. Benhouda, A.; Yahia, M. Toxicity and Anti-Inflammatory Effects of Methanolic Extract of *Umbilicus rupestris* L. Leaves (Crassulaceae). *Int J Pharma Bio Sci* **2015**, *6*, 395–408.
51. Bremner, P.; Rivera, D.; Calzado, M.A.; Obón, C.; Inocencio, C.; Beckwith, C.; Fiebich, B.L.; Muñoz, E.; Heinrich, M. Assessing Medicinal Plants from South-Eastern Spain for Potential Anti-Inflammatory Effects Targeting Nuclear Factor- κ B and Other pro-Inflammatory Mediators. *J Ethnopharmacol* **2009**, *124*, 295–305, doi:10.1016/j.jep.2009.04.035.
52. Hussain, M.S.; Azam, F.; Eldarrat, H.A.; Alskas, I.; Mayoof, J.A.; Dammona, J.M.; Ismail, H.; Ali, M.; Arif, M.; Haque, A. Anti-Inflammatory, Analgesic and Molecular Docking Studies of Lanostanoic Acid 3-O- α -D-Glycopyranoside Isolated from *Helichrysum stoechas*. *Arabian J Chem* **2020**, *13*, 9196–9206, doi:10.1016/j.arabjc.2020.11.004.
53. Hussain, M.S.; Azam, F.; Ahmed Eldarrat, H.; Haque, A.; Khalid, M.; Zaheen Hassan, M.; Ali, M.; Arif, M.; Ahmad, I.; Zaman, G.; Alabdallah, N.M.; Saeed, M. Structural, Functional, Molecular, and Biological Evaluation of Novel Triterpenoids Isolated from *Helichrysum stoechas* (L.) Moench. Collected from Mediterranean Sea Bank: Misurata- Libya. *Arabian J Chem* **2022**, *15*, 103818, doi:10.1016/j.arabjc.2022.103818.
54. Recio, M.C.; Giner, R.; Terencio, M.C.; Sanz, M.J.; Rios, J.L. Anti-Inflammatory Activity of *Helichrysum stoechas*. *Planta Med* **1991**, *57*, 365–371, doi:10.1055/s-2006-960317.
55. Noor, S.; Mohammad, T.; Rub, M.A.; Raza, A.; Azum, N.; Yadav, D.K.; Hassan, M.I.; Asiri, A.M. Biomedical Features and Therapeutic Potential of Rosmarinic Acid. *Arch Pharm Res* **2022**, *45*, 205–228, doi:10.1007/s12272-022-01378-2.
56. Bansal, Y.; Sethi, P.; Bansal, G. Coumarin: A Potential Nucleus for Anti-Inflammatory Molecules. *Med Chem Res* **2013**, *22*, 3049–3060, doi:10.1007/s00044-012-0321-6.
57. Pavlíková, N. Caffeic Acid and Diseases—Mechanisms of Action. *Int J Mol Sci* **2023**, *24*, 588, doi:10.3390/ijms24010588.

58. Aijaz, M.; Keserwani, N.; Yusuf, M.; Ansari, N.H.; Ushal, R.; Kalia, P. Chemical, Biological, and Pharmacological Prospects of Caffeic Acid. *Biointerface Res Appl Chem* **2023**, *13*, 324 doi:10.33263/BRIAC134.324.
59. Forouzanfar, F.; Sahranavard, T.; Tsatsakis, A.; Iranshahi, M.; Rezaee, R. Rutin: A Pain-Relieving Flavonoid. *Inflammopharmacology* **2025**, *33*, 1289–1301, doi:10.1007/s10787-025-01671-8.
60. Ginwala, R.; Bhavsar, R.; Chigbu, D.G.I.; Jain, P.; Khan, Z.K. Potential Role of Flavonoids in Treating Chronic Inflammatory Diseases with a Special Focus on the Anti-Inflammatory Activity of Apigenin. *Antioxidants* **2019**, *8*, 35, doi:10.3390/antiox8020035.
61. Proestos, C.; Lytoudi, K.; Mavromelanidou, O.K.; Zoumpoulakis, P.; Sinanoglou, V.J. Antioxidant Capacity of Selected Plant Extracts and Their Essential Oils. *Antioxidants* **2013**, *2*, 11–22, doi:10.3390/antiox2010011.
62. Krishnaiah, D.; Sarbatly, R.; Nithyanandam, R. A Review of the Antioxidant Potential of Medicinal Plant Species. *Food Bioprocess* **2011**, *89*, 217–233, doi:10.1016/j.fbp.2010.04.008.
63. Kasote, D.M.; Katyare, S.S.; Hegde, M. V.; Bae, H. Significance of Antioxidant Potential of Plants and Its Relevance to Therapeutic Applications. *Int J Biol Sci* **2015**, *11*, 982–991, doi:10.7150/ijbs.12096.
64. Amorati, R.; Valgimigli, L. Methods to Measure the Antioxidant Activity of Phytochemicals and Plant Extracts. *J Agric Food Chem* **2018**, *66*, 3324–3329, doi:10.1021/acs.jafc.8b01079.
65. Scherer, R.; Godoy, H.T. Antioxidant Activity Index (AAI) by the 2,2-Diphenyl-1-Picrylhydrazyl Method. *Food Chem* **2009**, *112*, 654–658, doi:10.1016/j.foodchem.2008.06.026.
66. Qa'dan, F.; Petereit, F.; Mansoor, K.; Nahrstedt, A. Antioxidant Oligomeric Proanthocyanidins from *Cistus salvifolius*. *Nat Prod Res* **2006**, *20*, 1216–1224, doi:10.1080/14786410600899225.
67. El Euch, S.K.; Cieřla, Ł.; Bouzouita, N. Free Radical Scavenging Fingerprints of Selected Aromatic and Medicinal Tunisian Plants Assessed by Means of TLC-DPPH Test and Image Processing. *J AOAC Int* **2014**, *97*, 1291–1298, doi:10.5740/jaoacint.SGEEI_Euch.
68. Balkan, B.; Balkan, S.; Aydođdu, H.; Güler, N.; Ersoy, H.; Ařkın, B. Evaluation of Antioxidant Activities and Antifungal Activity of Different Plants Species Against Pink Mold Rot-Causing *Trichothecium Roseum*. *Arab J Sci Eng* **2017**, *42*, 2279–2289, doi:10.1007/s13369-017-2484-4.
69. Georgieva, L.; Mihaylova, D. Screening of Total Phenolic Content and Radical Scavenging Capacity of Bulgarian Plant Species. *Int Food Res J* **2015**, *22*, 240–245.
70. Nassar-Eddin, G.; Zheleva-Dimitrova, D.; Danchev, N.; Vitanska-Simeonova, R. Antioxidant and Enzyme-Inhibiting Activity of Lyophilized Extract from *Clinopodium vulgare* L. (Lamiaceae). *Pharmacia* **2021**, *68*, 259–263, doi:10.3897/pharmacia.68.e61911.
71. Sarikurkcı, C.; Ozer, M.S.; Tepe, B.; Dilek, E.; Ceylan, O. Phenolic Composition, Antioxidant and Enzyme Inhibitory Activities of Acetone, Methanol and Water Extracts of *Clinopodium vulgare* L. Subsp. *Vulgare* L. *Ind Crops Prod* **2015**, *76*, 961–966, doi:10.1016/j.indcrop.2015.08.011.
72. Iyda, J.H.; Fernandes, Â.; Calhella, R.C.; Alves, M.J.; Ferreira, F.D.; Barros, L.; Amaral, J.S.; Ferreira, I.C.F.R. Nutritional Composition and Bioactivity of *Umbilicus rupestris* (Salisb.) Dandy: An Underexploited Edible Wild Plant. *Food Chem* **2019**, *295*, 341–349, doi:10.1016/j.foodchem.2019.05.139.
73. Wang, W.; Le, T.; Wang, W.W.; Yin, J.F.; Jiang, H.Y. The Effects of Structure and Oxidative Polymerization on Antioxidant Activity of Catechins and Polymers. *Foods* **2023**, *12*, 4207, doi:10.3390/foods12234207.
74. Wang, W.; Yue, R.F.; Jin, Z.; He, L.M.; Shen, R.; Du, D.; Tang, Y.Z. Efficiency Comparison of Apigenin-7-O-Glucoside and Trolox in Antioxidative Stress and Anti-Inflammatory Properties. *J Pharm Pharmacol* **2020**, *72*, 1645–1656, doi:10.1111/jphp.13347.
75. Álvarez-Martínez, F.J.; Barrajón-Catalán, E.; Herranz-López, M.; Micol, V. Antibacterial Plant Compounds, Extracts and Essential Oils: An Updated Review on Their Effects and Putative Mechanisms of Action. *Phytomedicine* **2021**, *90*, 153626, doi:10.1016/j.phymed.2021.153626.
76. Angelini, P. Plant-Derived Antimicrobials and Their Crucial Role in Combating Antimicrobial Resistance. *Antibiotics* **2024**, *13*, 746, doi:10.3390/antibiotics13080746.
77. Oulahal, N.; Degraeve, P. Phenolic-Rich Plant Extracts With Antimicrobial Activity: An Alternative to Food Preservatives and Biocides? *Front Microbiol* **2022**, *12*, 753518, doi:10.3389/fmicb.2021.753518.
78. Salam, M.A.; Al-Amin, M.Y.; Salam, M.T.; Pawar, J.S.; Akhter, N.; Rabaan, A.A.; Alqumber, M.A.A. Antimicrobial Resistance: A Growing Serious Threat for Global Public Health. *Healthcare* **2023**, *11*, 1946, doi:10.3390/healthcare11131946.

79. Zouine, N.; Ghachtouli, N. El; Abed, S. El; Koraichi, S.I. A Comprehensive Review on Medicinal Plant Extracts as Antibacterial Agents: Factors, Mechanism Insights and Future Prospects. *Sci Afr* **2024**, *26*, e02395, doi:10.1016/j.sciaf.2024.e02395.
80. Sánchez-Hernández, E.; Álvarez-Martínez, J.; González-García, V.; Casanova-Gascón, J.; Martín-Gil, J.; Martín-Ramos, P. *Helichrysum stoechas* (L.) Moench Inflorescence Extract for Tomato Disease Management. *Molecules* **2023**, *28*, 5861, doi:10.3390/molecules28155861.
81. Ríos, J.L.; Reccio, M.C.; Villar, A. Antimicrobial Activities of *Helichrysum stoechas*. *Planta Med* **1990**, *56*.
82. Kutluk, I.; Aslan, M.; Orhan, I.E.; Özçelik, B. Antibacterial, Antifungal and Antiviral Bioactivities of Selected *Helichrysum* Species. *S Afr J Bot* **2018**, *119*, 252–257, doi:10.1016/j.sajb.2018.09.009.
83. Albayrak, S.; Aksoy, A.; Sagdic, O.; Hamzaoglu, E. Compositions, Antioxidant and Antimicrobial Activities of *Helichrysum* (Asteraceae) Species Collected from Turkey. *Food Chem* **2010**, *119*, 114–122, doi:10.1016/j.foodchem.2009.06.003.
84. Bogdadi, H.A.A.; Kokoska, L.; Havlik, J.; Kloucek, P.; Rada, V.; Vorisek, K. *In Vitro* Antimicrobial Activity of Some Libyan Medicinal Plant Extracts. *Pharm Biol* **2007**, *45*, 386–391, doi:10.1080/13880200701215026.
85. Bayoub, K.; Baibai, T.; Mountassif, D.; Retmane, A.; Soukri, A. Antibacterial Activities of the Crude Ethanol Extracts of Medicinal Plants against *Listeria monocytogenes* and Some Other Pathogenic Strains. *Afr J Biotechnol* **2010**, *9*, 4251–4258.
86. Zalegh, I.; Bourhia, M.; Zerouali, K.; Katfy, K.; Nayme, K.; Khallouki, F.; Benzaarate, I.; Mohammad Salamatullah, A.; Alzahrani, A.; Nafidi, H.A.; Akssira, M.; Mhand, R.A. Molecular Characterization of Gene-Mediated Resistance and Susceptibility of ESKAPE Clinical Isolates to *Cistus monspeliensis* L. and *Cistus salviifolius* L. Extracts. *Evidence-Based Complementary Altern Med* **2022**, *2022*, 7467279, doi:10.1155/2022/7467279.
87. Alam, M.; Ahmed, S.; Elsbali, A.M.; Adnan, M.; Alam, S.; Hassan, M.I.; Pasupuleti, V.R. Therapeutic Implications of Caffeic Acid in Cancer and Neurological Diseases. *Front Oncol* **2022**, *12*, 860508, doi:10.3389/fonc.2022.860508.
88. Khan, F.; Bamunuarachchi, N.I.; Tabassum, N.; Kim, Y.M. Caffeic Acid and Its Derivatives: Antimicrobial Drugs toward Microbial Pathogens. *J Agric Food Chem* **2021**, *69*, 2979–3004, doi:10.1021/acs.jafc.0c07579.
89. Barroso, M.R.; Barros, L.; Dueñas, M.; Carvalho, A.M.; Santos-Buelga, C.; Fernandes, I.P.; Barreiro, M.F.; Ferreira, I.C.F.R. Exploring the Antioxidant Potential of *Helichrysum stoechas* (L.) Moench Phenolic Compounds for Cosmetic Applications: Chemical Characterization, Microencapsulation and Incorporation into a Moisturizer. *Ind Crops Prod* **2014**, *53*, 330–336, doi:10.1016/j.indcrop.2014.01.004.
90. Kherbache, A.; Senator, A.; Laouicha, S.; Al-Zoubi, R.M.; Bouriche, H. Phytochemical Analysis, Antioxidant and Anti-Inflammatory Activities of *Helichrysum stoechas* (L.) Moench Extracts. *Biocatal Agric Biotechnol* **2020**, *29*, 101826, doi:10.1016/j.bcab.2020.101826.
91. Carev, I.; Maravić, A.; Ilić, N.; Čilić, V.Č.; Politeo, O.; Zorić, Z.; Radan, M. UPLC-MS/MS Phytochemical Analysis of Two Croatian *Cistus* Species and Their Biological Activity. *Life* **2020**, *10*, 1–13, doi:10.3390/life10070112.
92. Gürbüz, P.; Demirezer, L.Ö.; Güvenalp, Z.; Kuruüzüm-Uz, A.; Kazaz, C. Isolation and Structure Elucidation of Uncommon Secondary Metabolites from *Cistus salviifolius* L. *Nat. Prod* **2015**, *9*, 175–183.
93. Jin, S.; Zhang, L.; Wang, L. Kaempferol, a Potential Neuroprotective Agent in Neurodegenerative Diseases: From Chemistry to Medicine. *Biomed. Pharmacother* **2023**, *165*, 115215, doi:10.1016/j.biopha.2023.115215.
94. Imran, M.; Salehi, B.; Sharifi-Rad, J.; Gondal, T.A.; Saeed, F.; Imran, A.; Shahbaz, M.; Fokou, P.V.T.; Arshad, M.U.; Khan, H.; Guerreiro, S.G.; Martins, N.; Estevinho, L.M. Kaempferol: A Key Emphasis to Its Anticancer Potential. *Molecules* **2019**, *24*, 2277, doi:10.3390/molecules24122277.
95. Periferakis, A.; Periferakis, K.; Badarau, I.A.; Petran, E.M.; Popa, D.C.; Caruntu, A.; Costache, R.S.; Scheau, C.; Caruntu, C.; Costache, D.O. Kaempferol: Antimicrobial Properties, Sources, Clinical, and Traditional Applications. *Int J Mol Sci* **2022**, *23*, 15054, doi:10.3390/ijms232315054.
96. Bangar, S.P.; Chaudhary, V.; Sharma, N.; Bansal, V.; Ozogul, F.; Lorenzo, J.M. Kaempferol: A Flavonoid with Wider Biological Activities and Its Applications. *Crit Rev Food Sci Nutr* **2023**, *63*, 9580–9604, doi:10.1080/10408398.2022.2067121.

97. Liu, M.-H.; Otsuka, N.; Noyori, K.; Shiota, S.; Ogawa, W.; Kuroda, T.; Hatano, T.; Tsuchiya, T. Synergistic Effect of Kaempferol Glycosides Purified from *Laurus nobilis* and Fluoroquinolones on Methicillin-Resistant *Staphylococcus aureus*. *Biol. Pharm. Bull.* **2009**, *32*, 489–492, doi:10.1248/bpb.32.489. 770–772
98. Huang, Y.H.; Huang, C.C.; Chen, C.C.; Yang, K.J.; Huang, C.Y. Inhibition of *Staphylococcus aureus* PriA Helicase by Flavonol Kaempferol. *Protein J* **2015**, *34*, 169–172, doi:10.1007/s10930-015-9609-y. 773–774
99. Fiamegos, Y.C.; Kastritis, P.L.; Exarchou, V.; Han, H.; Bonvin, A.M.J.J.; Vervoort, J.; Lewis, K.; Hamblin, M.R.; Tegos, G.P. Antimicrobial and Efflux Pump Inhibitory Activity of Caffeoylquinic Acids from *Artemisia absinthium* against Gram-Positive Pathogenic Bacteria. *PLoS One* **2011**, *6*, e18127, doi:10.1371/journal.pone.0018127. 775–777
100. Tian, J.; Fu, L.; Zhang, Z.; Dong, X.; Xu, D.; Mao, Z.; Liu, Y.; Lai, D.; Zhou, L. Dibenzo- α -Pyrones from the Endophytic Fungus *Alternaria* sp. Samif01: Isolation, Structure Elucidation, and Their Antibacterial and Antioxidant Activities. *Nat Prod Res* **2017**, *31*, 387–396, doi:10.1080/14786419.2016.1205052. 778–780
101. Bullitta, S.; Piluzza, G.; Manunta, M.D.I. Cell-Based and Chemical Assays of the Ability to Modulate the Production of Intracellular Reactive Oxygen Species of Eleven Mediterranean Plant Species Related to Ethnobotanic Traditions. *Genet Resour Crop Evol* **2013**, *60*, 403–412, doi:10.1007/s10722-012-9842-6. 781–783
102. Luís, A.; Neiva, D.; Pereira, H.; Gominho, J.; Domingues, F.; Duarte, A.P. Stumps of *Eucalyptus globulus* as a Source of Antioxidant and Antimicrobial Polyphenols. *Molecules* **2014**, *19*, 16428–16446, doi:10.3390/molecules191016428. 784–785
103. Gonçalves, J.; Luís, Â.; Gradillas, A.; García, A.; Restolho, J.; Fernández, N.; Domingues, F.; Gallardo, E.; Duarte, A.P. Ayahuasca Beverages: Phytochemical Analysis and Biological Properties. *Antibiotics* **2020**, *9*, 1–21, doi:10.3390/antibiotics9110731. 786–787
104. Coimbra, A.T.; Luís, Â.F.S.; Batista, M.T.; Ferreira, S.M.P.; Duarte, A.P.C. Phytochemical Characterization, Bioactivities Evaluation and Synergistic Effect of *Arbutus unedo* and *Crataegus monogyna* Extracts with Amphotericin B. *Curr Microbiol* **2020**, *77*, 2143–2154, doi:10.1007/s00284-020-02125-w. 788–790
105. Coimbra, A.; Miguel, S.; Ribeiro, M.; Coutinho, P.; Silva, L.A.; Ferreira, S.; Duarte, A.P. Chemical Composition, Antioxidant, and Antimicrobial Activities of Six Commercial Essential Oils. *Lett Appl Microbiol* **2023**, *76*, 1–8, doi:10.1093/lambio/ovac042. 791–792

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content. 793–795

Uma abordagem integrada à cultura de plantas silvestre e ao turismo em regiões de montanha por via de um sistema previsional inteligente de suporte à decisão em sustentabilidade: Montanha Viva

Montanha Viva – An intelligent prediction system for decision support in sustainability: An integrated approach in mountain regions to wild plants culture and tourism

Pedro Dinis Gaspar^{1,2}, Tânia Miranda Lima^{1,2}, José Pombo^{1,3}, Ana Paula Duarte^{1,4}, Jorge Monteiro⁵, Susana Ferreira^{1,4}, Angelo Luis^{1,4}, José Carlos Gonçalves^{6,7,8}, Pedro Neto⁹, Kelly O'Hara¹, Rui Brás¹, Sofia Santos¹⁰

⁽¹⁾ Universidade da Beira Interior, Rua Marquês D'Ávila e Bolama, 6201-001 Covilhã

⁽²⁾ C-MAST, Center for Mechanical and Aerospace Science and Technologies, Faculty of Engineering, University of Beira Interior, 6201-001 Covilhã.

⁽³⁾ Instituto de Telecomunicações, Department of Computer Science, University of Beira Interior, 6201-001 Covilhã, Portugal

⁽⁴⁾ Health Sciences Research Centre (CICS), University of Beira Interior, 6200-506 Covilhã, Portugal

⁽⁵⁾ Spaceway, LDA, R Pedro Nunes S/N, 3030-199, Santo Antonio Olivais, Coimbra

⁽⁶⁾ Plant Biotechnology Centre of Beira Interior (CBPBI), 6001-909 Castelo Branco, Portugal

⁽⁷⁾ Polytechnic Institute of Castelo Branco-School of Agriculture (IPCB-ESA), 6001-909 Castelo Branco, Portugal

⁽⁸⁾ Research Centre for Natural Resources, Environment and Society, Polytechnic Institute of Castelo Branco (CERNAS-IPCB), 6001-909 Castelo Branco, Portugal

⁽⁹⁾ Município do Fundão, Praça do Município 6230-338 Fundão Portugal

⁽¹⁰⁾ Gardunha 21, Praça do Município 6230-338 Fundão Portugal

Abstract: *Com vista a estimular, de forma integrada, o aproveitamento económico da cultura de plantas silvestres de regiões de montanha e o potencial turismo sustentável de montanha que lhe possa estar inerente, é desenvolvido um sistema inteligente de apoio à decisão em tempo real, especialmente projetado para a operação em localizações inóspitas e remotas (sem ligação à internet). Este Sistema de apoio à decisão visa contribuir para a redução de consumo de recursos naturais, promoção da biodiversidade e a preservação da sustentabilidade ambiental. Para tal, é realizada a identificação e caracterização de plantas de montanha com características potenciadoras de mitigação natural de pragas e doenças em culturas agrícolas e com propriedades de aplicação em saúde e bem-estar. É desenvolvido um sistema de sensorização local e remota por visão computacional. A análise do vigor das plantas recorre a algoritmos de inteligência artificial, que irá suportar a decisão na realização de atividades culturais em plantas existentes ou em novas explorações agroflorestais. Paralelamente, a análise do vigor das plantas ao longo do ano possibilita a adequação dinâmica de trajetos pedonais disponíveis na aplicação de turismo sustentável, no sentido de fomentar ou restringir a passagem em pontos específicos do trajeto*

dependendo do estado de crescimento da flora e da fauna.

Keywords: Intelligent and real-time decision support system, Mountain plants, Remote locations, Biodiversity, Environmental sustainability, Pests and diseases, Agricultural crops, Health and welfare, Artificial intelligence, Sustainable tourism.

Theme: Sustentabilidade / Tecnologias novas e emergentes / Tecnologias de Informação e Comunicação

Presentation: Oral

Advancing Smart Farming and Ecological Monitoring: Gathering Sensing, Computational Vision, Communications Technologies and Artificial Intelligence

Matilde Sousa ^{1,2}, Ana Alves ^{1,2}, Rodrigo Antunes ^{1,2}, Martim Aguiar ^{1,2}, Pedro Dinis Gaspar ^{1,2}, Nuno Pereira ^{1,2}

¹ C-MAST - Centre for Mechanical and Aerospace Science and Technologies, Covilhã, Portugal
galvao.sousa@ubi.pt (M.S.); ana.cristina.alves@ubi.pt (A.A.); rodrigo.antunes@ubi.pt (R.A.); mar-tim.aguiar@ubi.pt (M.A.); dinis@ubi.pt (P.D.G.); nuno.pereira@ubi.pt (N.P)

² Department of Electromechanical Engineering, University of Beira Interior, Calçada Fonte do Lameiro, 6200-358, Covilhã, Portugal;

* Correspondence: dinis@ubi.pt

Abstract

This study presents an in-depth exploration of the LITecS station, an innovative contribution to precision agriculture (PA), focusing on sustainable and biodiversity monitoring. Amidst the challenges of global population growth and the need for sustainable, high-yield agricultural practices, PA, supported by modern technology and data-driven methodologies, emerges as a pivotal approach for optimizing crop yield and resource management.

The LITecS station exemplifies the integration of Wireless Sensor Networks (WSN) into PA, offering real-time data collection on various environmental parameters. This system, characterized by its modular pole design, is self-sufficient, harnessing solar energy for continuous operation. It can be equipped with a diverse range of sensors to capture environmental and agricultural data. Additionally, the system's capability extends to real-time data transmission via LTE internet and/or LoRaWAN networks, ensuring timely and accurate delivery of information for cloud-based analysis.

Incorporating advanced sensor technology and sustainable energy solutions, the LITecS station addresses the critical challenges in PA. The system's design reflects a commitment to ecological stewardship and technological innovation, offering a model for future developments in sustainable and intelligent farming.

In conclusion, the LITecS station represents a significant advancement in PA, demonstrating the seamless integration of technology and ecological consciousness in modern agricultural management.

Keywords: Precision Agriculture, Wireless Sensor Networks, Internet of Things, Artificial Intelligence in Farming, Energy Efficiency in Agriculture, Sensor Integration, Smart Farming Technologies.

Academic Editor: Firstname Last-name

Received: date

Revised: date

Accepted: date

Published: date

Citation: To be added by editorial staff during production.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Agriculture and farming, is vital for sustenance and human life, and have witnessed a transformative shift in the wake of technological advancements. With the increase of global population, the demand for sustainable, high-yield agricultural practices has

become imperative [1]. Precision Agriculture (PA), driven by modern technology and data-driven methodologies, emerges as a pivotal approach to optimize crop yield and resource management [2].

The integration of Wireless Sensor Networks (WSNs) into PA holds immense potential. WSNs enable real-time data collection on various environmental parameters, providing farmers a comprehensive understanding of their crops' needs [3]. These networks serve multifaceted purposes, from early pest and disease detection to climate monitoring, and even facilitating precise management techniques like targeted pesticide application [4].

However, the effective utilization of WSNs in agriculture requires addressing crucial challenges, particularly concerning energy efficiency and routing protocols. Energy-efficient routing schemes are crucial to extend the lifespan of sensor networks, ensuring continuous and reliable data transmission [5,6].

Precision agriculture, a contemporary agricultural trend, seeks to optimize production efficiency while minimizing environmental impacts [7]. Its applications extend to epidemic disease control, mitigating the adverse effects of climate-induced diseases and reducing the excessive use of chemical fungicides [8].

The convergence of Internet of Things (IoT) technologies with PA improves distributed monitoring and control systems, revolutionizing diverse application areas [9]. However, challenges persist, particularly concerning energy management and interoperability issues within heterogeneous sensor networks [10].

Efforts to standardize communication interfaces and establish consensus-based standards are key to simplify the integration of diverse sensors, reducing complexities and boost interoperability in data acquisition networks [11,12].

In essence, the combination of PA, IoT technologies, and WSN presents a tchange in the landscape of agricultural practices. Addressing energy efficiency and standardization challenges is crucial to unleash the full potential of these advancements in agricultural sustainability and productivity.

In this context, this paper presents a detailed examination of two innovative systems, the Montanha Viva system and the LITecS station, as case studies in the PA sector. These systems represent cutting-edge solutions that integrate advanced sensor technology with sustainable energy solutions to address the evolving challenges faced by modern agriculture. Through the analysis of these case studies, this paper aims to elucidate the practical applications, technological features, and benefits of such integrated systems in improving agricultural resilience, biodiversity monitoring, and environmental sustainability.

2. Literature Review

The fusion of IoT, WSNs, and artificial intelligence (AI) is revolutionizing PA, addressing critical challenges and leveraging new opportunities for sustainable farming practices. This comprehensive review covers key aspects of modular monitoring systems, IoT-based disease forecasting, energy optimization in WSNs, and advanced sensor network management frameworks.

Khattab et al. [9], demonstrates IoT pivotal roles in early epidemic disease detection in crops. The system stores environmental and soil information from wireless sensor networks, allowing real-time monitoring through internet-enabled devices. Utilizing sensors, microcontrollers, and AI-driven data fusion, IoT frameworks have successfully detected diseases in tomatoes and potatoes, showcasing its transformative impact on agricultural decision-making and disease control. Field experiments demonstrate that the system minimizes chemical usage, enhances crop quality, and is versatile for various plant disease models. Figure 1 displays the agricultural weather station used, including a detailed view of its controller box.

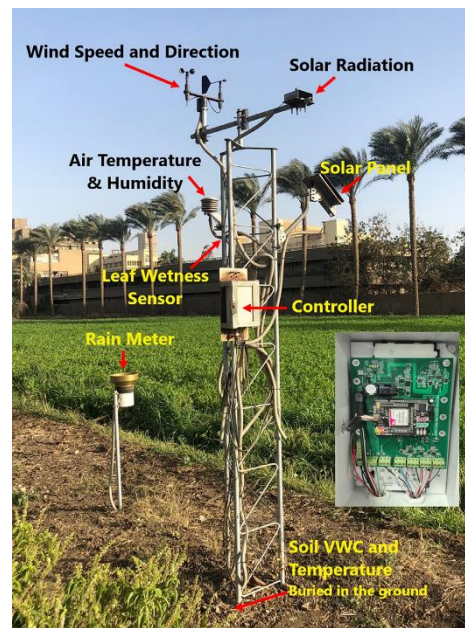


Fig. 1. The agro-weather station used by Khattab et al. [9].

Fernandes et al. [13] discusses the interoperability problem in wireless sensor networks for PA and Precision Viticulture (PV). Research highlights the implementation of IEEE 1451-based frameworks in managing diverse sensors within PA. These frameworks address challenges in sensor integration and streamline data acquisition, enhancing precision and efficiency in agricultural data management. Furthermore, the article provides valuable insights into the effective implementation of IEEE 1451, presenting practical guidance for the development of intelligent and efficient PA/PV technologies. Figure 2, represent the architecture of an IEEE 1451 WSN compliant network.

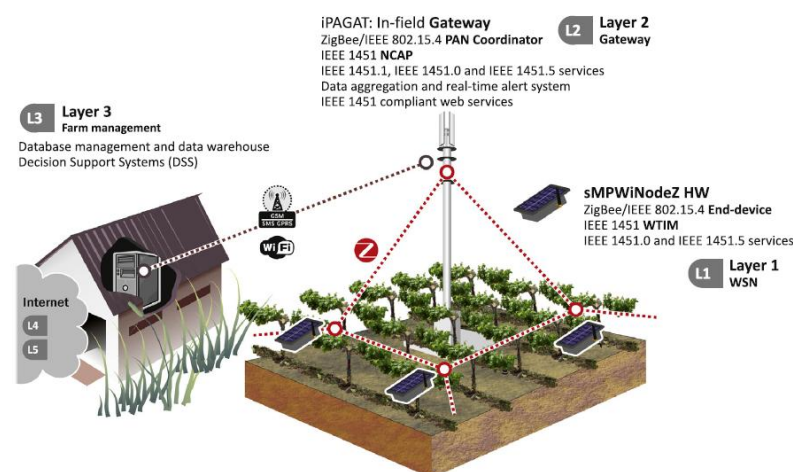


Fig. 2. Illustration of an IEEE 1451 WSN compliant network deployed by Fernandes et al., [13] over a vineyard, emphasizing IEEE 1451 concepts and entities. Each WSN node is a WTIM device while the NCAP acts as a sink node (gateway) to collected data.

Pandiyaraju et al. [14] focuses on improving WSNs in PA by optimizing energy consumption in sensor nodes. It introduces a multi-objective clustering approach, a hybrid optimization technique, and a Convolutional Neural Network (CNN) integration. The approach improves performance metrics like classification accuracy, throughput, packet delivery ratio, network lifetime, and energy consumption. Potential enhancements include

a new routing mechanism, a lightweight optimizer, and a lightweight CNN to further refine the approach and improve crop yield.

Focusing on energy efficiency, recent innovations in WSNs employ multi-objective clustering methods and deep learning techniques to optimize energy usage. These methods enhance classification accuracy, extend network lifetimes, and reduce energy consumption, offering robust solutions to energy challenges in PA. Haseeb et al. [15], proposed an IoT-based WSN framework for smart agriculture, utilizing agricultural sensors to capture data and determine cluster heads. The framework measures signal strength and security, resulting in improved communication performance. Simulations show a 13.5% increase in network throughput, 38.5% packets drop ratio, 13.5% network latency, 16% energy consumption, and 26% routing overheads compared to other solutions. Figure 3 depicts a smart agriculture scenario utilizing sensors, sink nodes, BS, Internet, and users.

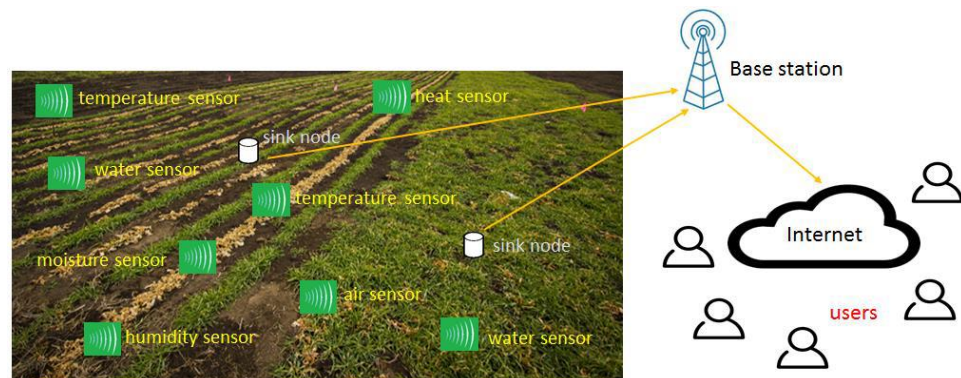


Fig. 3. Smart agricultural environment based on wireless sensor network (WSN) used by Haseeb et al. [15].

Exploring genetic algorithm-based routing protocols, Patil & Kohle [16], studies emphasize the importance of high-scalability and low-latency solutions for improving WSN performance, particularly in time-sensitive applications. This protocol was suitable for highly distributed and rapidly expanding networks, outperforming conventional routing algorithms and achieving lower latency in highly scalable WSNs.

Research on low-cost WSNs for vine phenology monitoring, especially flowering stages. The DHT22 sensor, chosen for its low cost and high accuracy, was integrated into a network with a central node and several remote nodes connected via Zigbee communication protocols. This innovative approach highlights the potential for cost-effective, accurate, and detailed measurement of spatial variability in agriculture [17]. Figure 4 depicts the arrangement of the spatialized sensors along with the weather station.

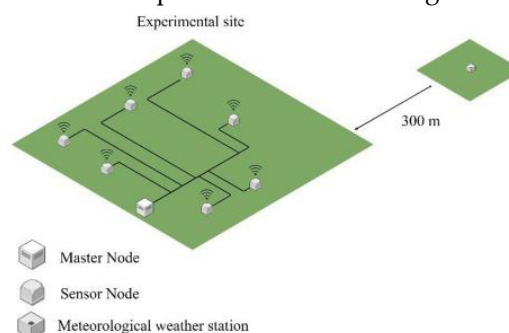


Fig.4. Illustration of the distribution of the spatialized sensors and the weather station used by Fuentes-Peñailillo et al. [17].

Dehwah et al. [18], discuss the Universal Dynamic Weather-Conditioned Moving Average (UD-WCMA) as a new energy estimation and forecasting algorithm for solar-powered wireless sensor networks. UD-WCMA uses real-time data and historical energy patterns to improve prediction accuracy in diverse weather conditions. The system used custom-developed sensor boards powered by a 32-bit ARM Cortex M4 micro-controller and a 20Wp/20.8Voc solar panel.

Zhang et al. [19], focused on organic photovoltaic energy harvesting systems that have demonstrated the viability of OPV(Organic Photovoltaic)-powered WSNs in indoor settings, contributing to sustainable energy solutions for sensor networks. The system used sensors for temperature, humidity, air pressure, CO₂ levels, real-time clock, and battery voltage. The sensor nodes were powered by a LiPo battery and a flexible OPP (Organic Photovoltaic Panel) module from InfinityPV Denmark. Data transmission was done using a ZigBee RF module for ease of deployment and high data rate.

Foughali et al. [20], investigated into a cost-effective WSN for potato farming, utilizing the NodeMCU IoT platform and DHT11 sensor for temperature and humidity measurements was developed. The system, powered by two 1.5V batteries and housed in IP66-rated enclosures, was designed to monitor micro-climate conditions, crucial for managing and preventing late blight in potato crops. This setup allows for efficient and cost-effective monitoring of micro-climate conditions, which is crucial for managing and preventing late blight in potato crops.

These studies collectively emphasize the critical role of IoT, WSN, and AI in advancing PA. They highlight the importance of sensor integration, energy efficiency, AI-based forecasting, and cost-effective solutions in shaping future agricultural practices that are data-driven, efficient, and environmentally sustainable. The integration of these technologies facilitates the development of modular monitoring systems that are adaptable, scalable, and efficient, catering to the evolving needs of modern agriculture [21].

3. Materials and Methods

This section presents a study of the LITecS station, an innovative contribution to Precision agriculture with a focus on sustainable monitoring.

The LITecS station exemplifies the integration of Wireless Sensor Networks (WSNs) into PA, offering real-time data collection on multiple environmental parameters. The system is characterized by its modular pole design, which confers it modularity. It is self-sufficient, harnessing solar energy for continuous operation, and its primary advantage is the ability to be equipped with customized sensors, selected to meet specific requirements. Depending on the objective, sensors for soil moisture, temperature, air humidity, or rain and wind can be integrated. Additionally, the system is capable of real-time data transmission via LTE and/or LoRaWAN networks, ensuring the timely delivery of information for cloud-based analysis.

By incorporating advanced sensor technology and sustainable energy solutions, the LITecS station addresses the critical challenges in PA, including energy efficiency, routing protocols, and the seamless integration of diverse sensors. The system's design reflects a commitment of improving ecological practices technological innovation, offering a model for future developments in intelligent and sustainable agriculture.

The sensor configuration of the LITecS station is adaptable. As an example, a possible configuration may consist of a thermometer, a pluviometer, an anemometer, a wind direction sensor, a barometer, and a hygrometer, also allowing for the integration of cameras. However, as the system is modular, it is simple to remove, replace, or add new sensors and cameras, thus adapting it to different monitoring scenarios. Figure 5 illustrates one such configuration example.

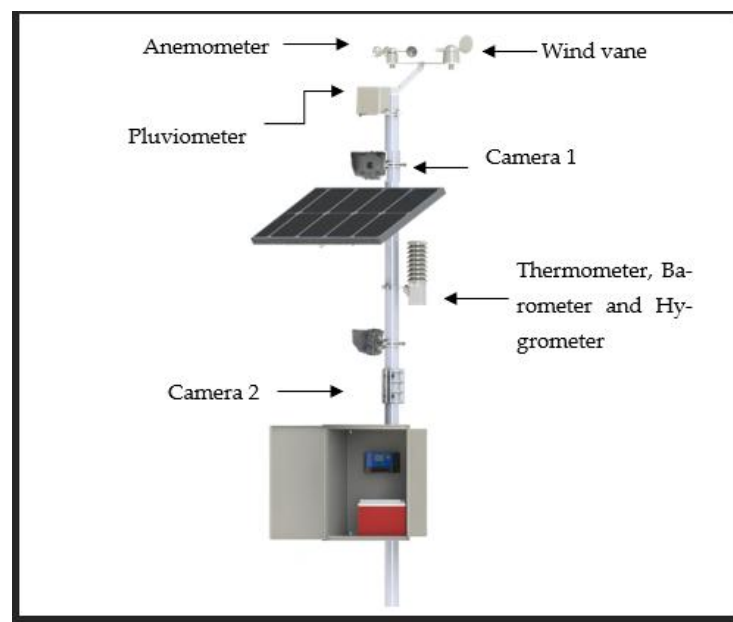


Fig. 5. Example of sensors used in the LITecS station, the station is modular, so it allows the integration of multiple sensors and cameras.

The design and deployment of the LITecS Station are customized to meet the needs of real-time collection, processing, and visualization of environmental and agricultural data. Projected for precision, efficiency, and scalability, the system's architecture anticipates the complex requirements of environmental monitoring.

The data collection system consists of two main parts and utilizes cloud computing to reduce costs and improve energy efficiency. The station is composed of cameras to capture images of the monitored flora, along with a set of environmental sensors selected for the specific application to measure essential parameters.

The station's sensors can record temperature, humidity, pressure, soil moisture, precipitation, mean wind speed, maximum wind speed, and wind direction. Measurements are taken by an ESP32 every 10 minutes. This interval was chosen in line with World Meteorological Organization (WMO) [1] recommendations for calculating mean wind speed in meteorological records. The ESP32 processes raw data to calculate the mean and maximum wind speeds and the total precipitation over the interval. The processed data are then sent to Raspberry Pi Zero, which compiles them into a CSV file and uploads it to a cloud storage system capable of storing, processing, and managing the data.

The use of cameras, in conjunction with computer vision systems, allows the analysis of a variety of environmental and agricultural factors. For instance, they can be used to consistently monitor the same area, monitoring the plant's phenology over time. Images are captured by a Raspberry Pi Zero 2W. An ESP32 controls the power to the Raspberry Pi, activating it only for the time necessary to capture and transmit the image. This timed operation conserves energy, supporting the objective of a more efficient system.

The data stored in the cloud are processed through a pipeline that handles the two data types in parallel. For the sensor data, the pipeline retrieves the files from the cloud and uploads them to a MySQL database. For the images, the pipeline performs pre-processing before applying a computer vision model to, for example, determine the plant's phenological stage. Both the processed image and the classification result are then stored in the MySQL database for display on the dashboard.

Data integration from both the sensor and imaging branches is managed by the system's back-end infrastructure, which retrieves processed data from the MySQL database.

The back-end, implemented using FastAPI, coordinates the handling of environmental measurements and phenology classifications. This configuration ensures that meteorological data and insights from plant phenology are combined to enable comprehensive monitoring of environmental conditions and the plant's different developmental stages.

The front-end of the system, developed with React, presents the collected information via an interactive dashboard. Users can explore real-time and historical meteorological data, view plant images with their processed status, and upload photos to be classified by the model. This design not only provides a clear and actionable view of the monitored ecosystem but also fosters a symbiotic relationship with users, who contribute valuable data to further enhance the model.

A summary of the LITecS station's architecture, including its hardware components and the data acquisition, processing, and visualization flow, is illustrated in Figure 6, and upload pictures to be classified by the model. This design not only provides a clear and actionable view of the monitored ecosystem but also fosters a symbiotic relationship with users, who contribute valuable data to further improve the model.

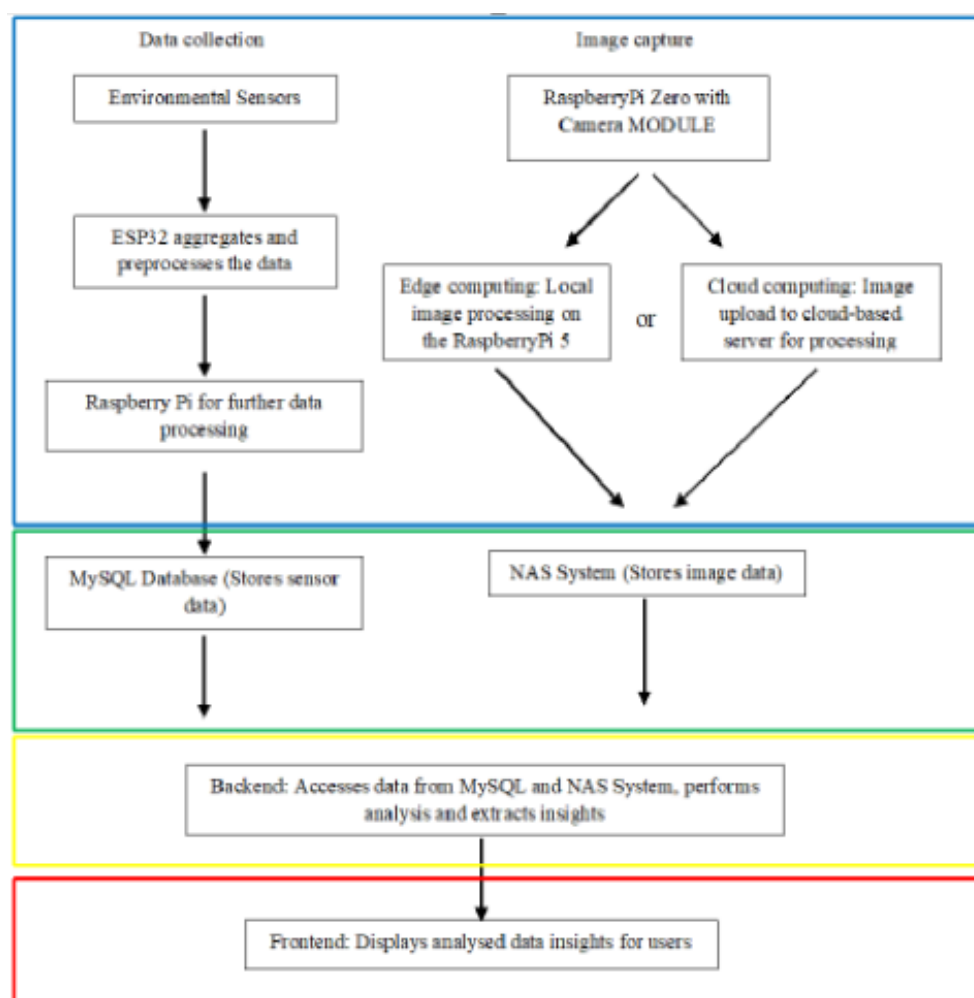


Fig .6. Flow diagram representing the sequence steps of the data flow architecture in the LITecS station: (blue) Data capture and processing; (green) Data transmission and storage; (yellow) Data analysis, machine learning for species identification; (red) web server and user interface

This research highlights the significant role of IoT, WSNs, and AI in advancing Precision Agriculture. It emphasizes their application in improving weather forecasts, water management, and strategies for extreme weather events, alongside contributions to climate research, agriculture, and environmental monitoring. It also underlines the

importance of flexibility in sensor integration, improving energy efficiency, and employing AI for forecasting to cultivate data-driven, efficient, and sustainable agricultural practices. The LITecS station exemplifies the fusion of technology with ecological sustainability in agriculture, showcasing how advanced monitoring capabilities empower agriculturists to make informed, environmentally responsible decisions, illustrating the effective combination of technological innovation and environmental responsibility in contemporary agricultural management.

4. Case Studies

4.1 BioD'Agro System Case Study

In the field of Precision Agriculture (PA), with a specific application in viticulture, the BioD'Agro system stands out by merging advanced sensor technology with green energy solutions to transform vineyard management and agricultural biodiversity monitoring. This system meets the needs of modern agricultural management, embodying both environmental practices and technological innovation.

Central to its operation is the LITecS station, which efficiently gathers a broad spectrum of data. This includes measurements of soil moisture, temperature, nutrient levels, wind speed and direction, and solar exposure. In addition to these, specialized sensors were integrated, such as those for wet leaf detection and a microphone for monitoring bats. Since bats are natural predators of pests that affect vineyards, tracking their activity makes it possible to predict and mitigate potential infestations. Real-time data transmission is facilitated through LTE networks. Furthermore, the integration of cameras enriches the dataset with imagery, enabling the application of computer vision techniques to assess parameters such as greenness levels, growth rates, infection detection, and moisture estimation [NP2].

In an era that emphasizes ecological responsibility, the BioD'Agro system illustrates how technology can foster more sustainable viticulture. It provides in-depth insights into vineyard conditions, empowering farmers and producers to make informed, preventative decisions. This advancement marks a significant leap in PA, showcasing a commitment to maintaining the balance of agricultural ecosystems.

4.2 Montanha Viva System Case Study

The LITecS stations were implemented as part of the Montanha Viva project, an initiative aimed at protecting and improving mountain areas and their native wild flora. By identifying, characterizing, and monitoring the phenology of plant species, the project promotes nature conservation while supporting local tourism and recreational activities.

As part of this project, three stations were installed at different locations on the slopes of the Serra da Gardunha. These stations consist of a thermometer, a pluviometer, an anemometer, a wind direction sensor, a barometer, a hygrometer, and two cameras. Data was collected daily for a year, covering all seasons and a variety of weather conditions. The primary application in this case is related to local tourism: based on the phenological state of the plants, analyzed from the captured images, the system can recommend the best hiking routes for users, thereby promoting ecotourism in the region. Concurrently, the diverse meteorological data collected by the sensors are also provided to local farmers to support their agricultural activities.

Figure 7 shows example images captured by each station, demonstrating the consistency and quality of the imaging system.



Fig. 7.-Examples of images retrieved by the imaging system

5.Results

This section presents the practical results obtained from the data collected by the LITecS stations in the different case studies.

5.1. Results BioD’Agro System Case Study

The images captured by the stations were processed using various deep learning models to extract relevant agronomic and environmental information. For semantic segmentation, the DeepLabv3 architecture was used. This model allowed for the identification and isolation of different elements in the images, such as soil and vegetation. This technique was fundamental to automatically analyzing the type of plants between the vineyard rows and assessing the leaf state, identifying areas with different vigor or potential water stress. From the segmentation, it was also possible to calculate the greenness level (greenness index), a key indicator of the plant's health and photosynthetic activity.



Fig. 8. Example of the model segmentation of the vineyard

The Autonomous Bat Echolocation (ABE) Monitoring System uses ultrasonic microphones to capture the echolocation vocalizations of bats. Acoustic data is processed in real-time by the BatDetect2 deep learning model, which detects and classifies bat species based on their acoustics. Because bat activity is correlated with insect abundance, the system functions as an early warning mechanism for potential pest infestations. Since bats are natural predators of insects that constitute pests for vineyards, their presence and activity serve as a bioindicator of ecosystem health. The model achieved a detection

accuracy of 0.94 (AP Det) and a mean species classification accuracy of 0.85 (mAP Class). To deepen this understanding, a analysis revealed that the diet of bats in the area included the "meadow spittlebug" (*Philaenus spumarius*), a known pest in the viticultural sector. Figure 9 presents a spectrogram of an audio recording, where the model marks the detection of the bat's acoustics, proving the activity of these predators to be correlated with the environmental conditions and the phenology of the vineyard.

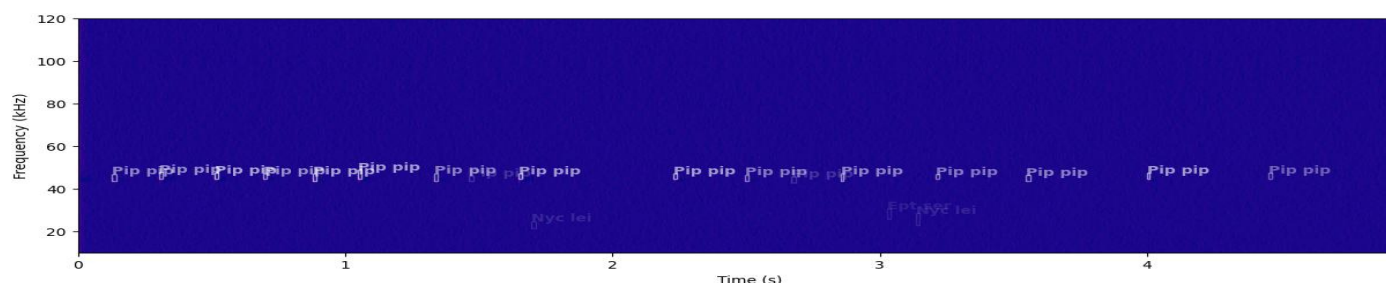


Fig. 9. Example of the spectrogram of an audio recording

To make the collected data accessible, an interactive web dashboard, as illustrated in Figure 10, was developed that centralizes all the information. The platform centralized all the information, allowing farmers and agricultural technicians to analyze sensor data through interactive tables and graphs, compare different variables, and export the results. A visual monitoring component, composed of a photo gallery, complemented the data analysis, offering farmers a remote and real-time view of the vegetative state of the vineyards.



Fig. 10. Representation of the Biod'Agro dashboard

5.2.Results Montanha Viva System Case Study

As part of the Montanha Viva project, a computer vision pipeline was implemented for the automated analysis of captured images. This process utilized the YOLOv8 model for the initial detection of plants in photographs, followed by the EfficientNetB5 model for the precise classification of the species and its respective phenological state. After this processing, the images were made available on the dashboard for user visualization. The extracted phenological information formed the basis for a route recommendation system, which suggested to users the hiking trails with the highest probability of observing certain species in interesting phases, such as flowering. Additionally, the platform offered robust tools for the analysis of sensor data: the user could access the data in table or graph format as presented in Figure 11, export the information in CSV files as well as the graphs themselves, and perform a comparative analysis between different parameters measured by the station's sensors. The user can also upload pictures to be classified by the model. This design not only provides a clear and actionable view of the monitored ecosystem but also fosters a symbiotic relationship with users, who contribute valuable data to further improve the model.

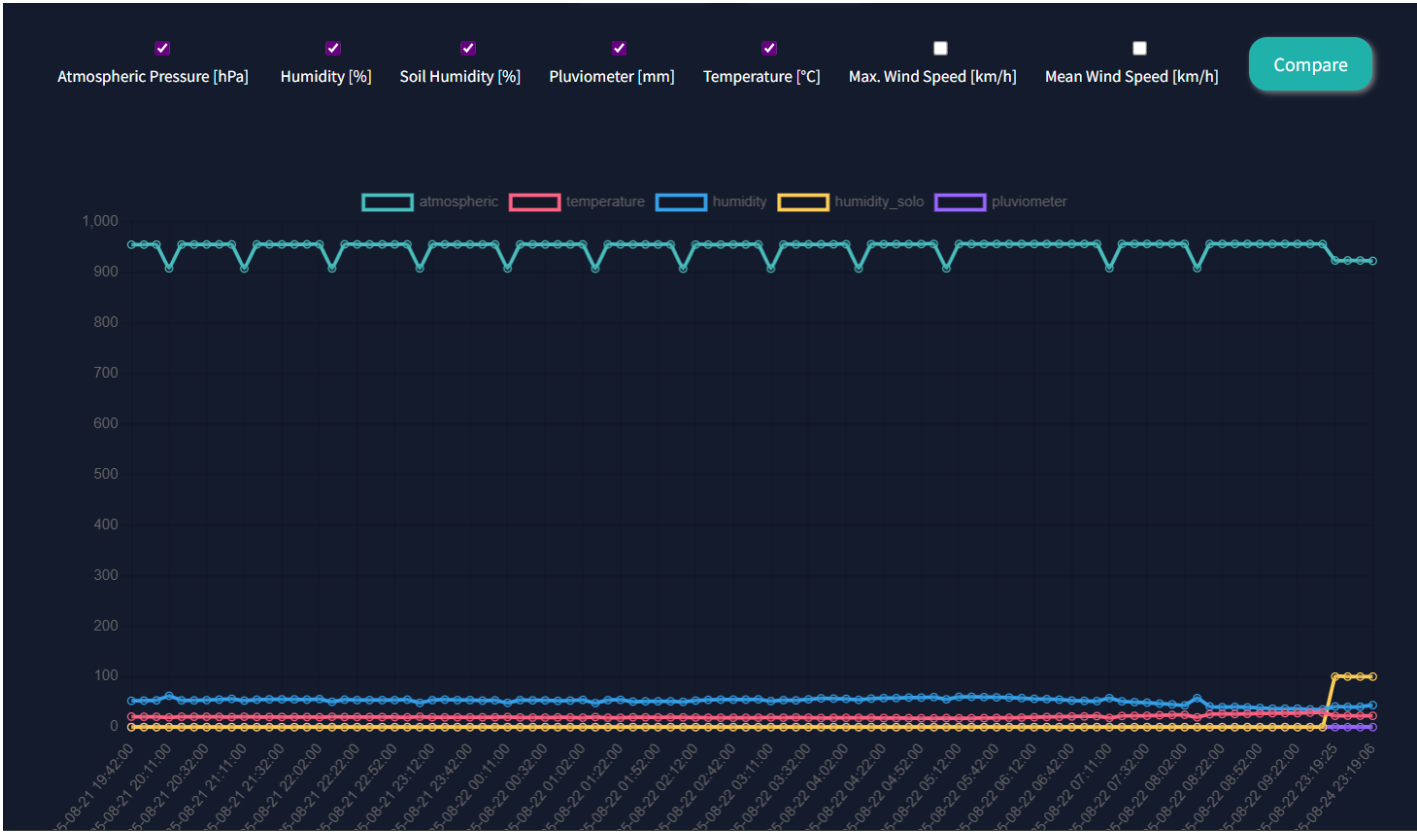


Fig. 11. Examples of the different graphs that user can access on the Montanha Viva Dashboard

6.Conclusions

The LITecS station and associated systems underscores a pivotal shift in Precision Agriculture (PA) through the strategic integration of sensor technology, IoT, and Wireless Sensor Networks (WSNs). The Montanha Viva and BioD'Agro case studies showcase the transformative impact of these technologies in addressing contemporary agricultural challenges, fostering sustainable farming practices, and enhancing biodiversity monitoring.

A key feature of the system presented is its modular design, enabling customizable sensor arrays to meet varied agricultural monitoring objectives efficiently. This flexibility, combined with real-time data acquisition and energy-efficient operations, represents a significant advancement in agricultural technologies, allowing for precise, data-driven management practices that align with environmental sustainability goals.

The LITecS station, central to the technological solutions explored, exemplifies the effective integration of ecological considerations with technological innovation within the agricultural sector. It marks a progressive step towards harmonizing agricultural productivity with the preservation of ecological balance, demonstrating the potential of modern technology to revolutionize 21st-century farming practices.

Looking ahead, the advancement of PA depends on continued innovation and the resolution of existing challenges, particularly in areas of energy efficiency and system interoperability. Collaborative research efforts are essential in refining these technologies, ensuring their adaptability and effectiveness in meeting the evolving needs of global agriculture. The continued development and application of such technologies hold the promise of a sustainable, productive agricultural future, reinforcing the indispensable role of computational and electronic advancements in the stewardship of our agricultural and natural ecosystems.

Acknowledgment

The authors acknowledge the support provided by LITecS (Laboratory of Innovation and Technologies for Sustainability) (www.litecs.ubi.pt).

Funding

This is within the activities of project Montanha Viva – An intelligent prediction system for decision support in sustainability, project PD21-00009, promoted by PROMOVE program funded by Fundação La Caixa and supported by Fundação para a Ciência e a Tecnologia and BPI.

This research was partially funded by the Portuguese Foundation for Science and Technology, I.P. (FCT, I.P.) FCT/MCTES through national funds (PIDDAC), under the R&D Unit C-MAST/Center for Mechanical and Aerospace Science and Technologies, reference: Projects UIDB/00151/2020 (<https://doi.org/10.54499/UIDB/00151/2020>) and UIDP/00151/2020 (<https://doi.org/10.54499/UIDP/00151/2020>)

References

1. Ahmed, N., De, D., & Hussain, I. (2018). Internet of Things (IoT) for smart precision agriculture and farming in rural areas. *IEEE internet of things journal*, 5(6), 4890–4899.
2. Popescu, D., Stoican, F., Stamatescu, G., Ichim, L., & Dragana, C. (2020). Advanced UAV–WSN system for intelligent monitoring in precision agriculture. *Sensors*, 20(3), 817.
3. Wang, S. (2021). Multipath routing based on genetic algorithm in wireless sensor networks. *Mathematical Problems in Engineering*, 2021(1), 4815711.
4. Akhter, R., & Sofi, S. A. (2022). Precision agriculture using IoT data analytics and machine learning. *Journal of King Saud University-Computer and Information Sciences*, 34(8), 5602–5618.
5. Nehra, V., Sharma, A. K., & Tripathi, R. K. (2019). NMR inspired energy efficient protocol for heterogeneous wireless sensor network. *Wireless Networks*, 25(6), 3689–3700.
6. Thangaramya, K., Kulothungan, K., Indira Gandhi, S., Selvi, M., Santhosh Kumar, S. V. N., & Arputharaj, K. (2020). RE-TRACTED ARTICLE: Intelligent fuzzy rule-based approach with outlier detection for secured routing in WSN. *Soft Computing*, 24(21), 16483–16497.
7. Rogers, D., & Tsirkunov, V. (2011). Costs and benefits of early warning systems. *Global assessment rep.*

8. Lim, J. A., Yaacob, J. S., Mohd Rasli, S. R. A., Eyahmalay, J. E., El Enshasy, H. A., & Zakaria, M. R. S. (2023). Mitigating the repercussions of climate change on diseases affecting important crop commodities in Southeast Asia, for food security and environmental sustainability—A review. *Frontiers in Sustainable Food Systems*, 6, 1030540.
9. Khattab, A., Habib, S. E., Ismail, H., Zayan, S., Fahmy, Y., & Khairy, M. M. (2019). An IoT-based cognitive monitoring system for early plant disease forecast. *Computers and Electronics in Agriculture*, 166, 105028.
10. Tani, F. K., & Cugnasca, C. E. (2005). Agriculture and the IEEE 1451 smart transducer interface standards. *EFITA WCCA 2005 proceedings*.
11. Chen, C., & Helal, S. (2008). Sifting through the jungle of sensor standards. *IEEE Pervasive Computing*, 7(4), 84-88.
12. Oostdyk, R. L., Mata, C. T., & Perotti, J. M. (2006, March). A Kennedy Space Center implementation of IEEE 1451 networked smart sensors and lessons learned. In *2006 IEEE Aerospace Conference* (pp. 20-pp). IEEE.
13. Fernandes, M. A., Matos, S. G., Peres, E., Cunha, C. R., López, J. A., Ferreira, P. J. S. G., ... & Morais, R. (2013). A framework for wireless sensor networks management for precision viticulture and agriculture based on IEEE 1451 standard. *Computers and Electronics in Agriculture*, 95, 19-30.
14. Pandiyaraju, V., Ganapathy, S., Mohith, N., & Kannan, A. (2023). An optimal energy utilization model for precision agriculture in WSNs using multi-objective clustering and deep learning. *Journal of King Saud University-Computer and Information Sciences*, 35(10), 101803.
15. Haseeb, K., Ud Din, I., Almogren, A., & Islam, N. (2020). An energy efficient and secure IoT-based WSN framework: An application to smart agriculture. *Sensors*, 20(7), 2081.
16. Patil, V. B., & Kohle, S. (2024). A high-scalability and low-latency cluster-based routing protocol in time-sensitive WSNs using genetic algorithm. *Measurement: Sensors*, 31, 100941.
17. Fuentes-Peñailillo, F., Acevedo-Opazo, C., Ortega-Farías, S., Rivera, M., & Verdugo-Vásquez, N. (2021). Spatialized system to monitor vine flowering: Towards a methodology based on a low-cost wireless sensor network. *Computers and Electronics in Agriculture*, 187, 106233.
18. Dehwah, A. H., Elmetennani, S., & Claudel, C. (2017). UD-WCMA: An energy estimation and forecast scheme for solar powered wireless sensor networks. *Journal of Network and Computer Applications*, 90, 17-25.
19. Zhang, S., Bristow, N., David, T. W., Elliott, F., O'Mahony, J., & Kettle, J. (2022). Development of an organic photovoltaic energy harvesting system for wireless sensor networks; application to autonomous building information management systems and optimisation of OPV module sizes for future applications. *Solar energy materials and solar cells*, 236, 1115503.
20. Foughali, K., Fathallah, K., & Frihida, A. (2018). Using Cloud IOT for disease prevention in precision agriculture. *Procedia computer science*, 130, 575-582.
21. Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, 9, 4843-4873.

Soluções Sustentáveis para o Setor Agroindustrial (Aviso n.º 02/SIAC/2019 - Ref. 46425)

Uma abordagem integrada à cultura de plantas silvestre e ao turismo em regiões de montanha por via de um sistema previsional inteligente de suporte à decisão em sustentabilidade: Montanha Viva

Pedro Dinis Gaspar, Tânia Lima, José Pombo, Ana Paula Duarte, Jorge Monteiro, Susana Ferreira, Ângelo Luís, José Carlos Gonçalves, Pedro Neto, Kelly O'Hara, Rui Brás and Sofia Santos

Universidade da Beira Interior, Portugal

dinis@ubi.pt

Apoiado pelo Programa Promove da Fundação “la Caixa”, em colaboração com o BPI e com a Fundação para a Ciência e a Tecnologia (FCT)

Contexto Global - Montanha e Plantas



Food and Agriculture
Organization of the
United Nations



Benefits of mountains



1

14% da
população mundial



2

25% da
biodiversidade



3

Fonte de 30% das
culturas mais
importantes



4

15% do turismo
global

22% da área
da Terra



5

70% da
água doce



Food and Agriculture
Organization of the
United Nations



Benefits of healthy plants



1

Saúde e
Alimentação



2

Fertilidade dos
Solos



3

Retenção da
Água



4

Mitigação de
Alterações
Climáticas



5

Protege
Biodiversidade



United Nations

O potencial da agricultura de montanha está ainda por explorar



Contexto Regional - Problemática

Serra da Gardunha - Região Ibérica de montanha do interior constituída por flora silvestre e agricultura de subsistência

- Desafios Ambientais
 - **Alterações climáticas**
(□ temperatura, □ água, □ pragas e doenças)
 - Pouca resiliência a catástrofes (que estão a aumentar)
 - Diminuição da Biodiversidade (flora silvestre)
- Desafios Logísticos
 - Aprovisionamento, distribuição e manutenção
 - **Acesso e retenção da água**
 - **Desertificação** e mão-de-obra
 - Baixos níveis de aproveitamento dos recursos naturais
- Desafios Tecnológicos
 - **Fraca** ou nenhuma ligação à **internet**
 - **“Resistência”** tecnológica das populações



É fundamental

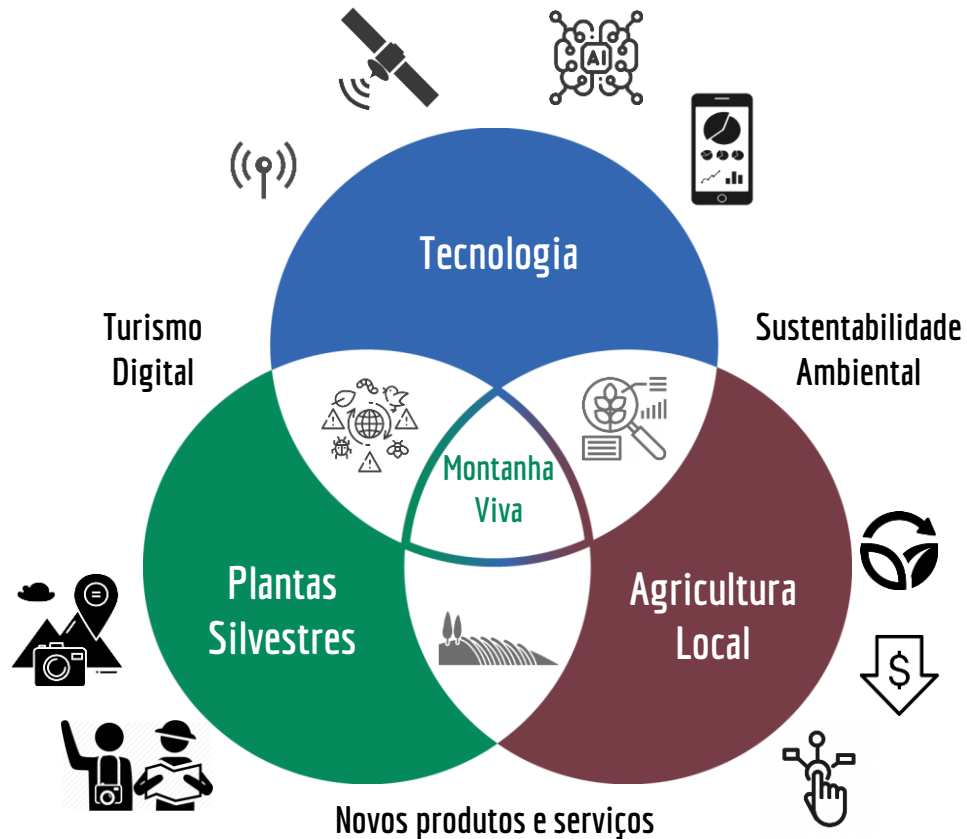
- Tornar as culturas agrícolas mais resilientes e sustentáveis.
 - **Diminuir o consumo de recursos naturais**, e em particular da água;
 - **Diminuir a aplicação de produtos agroquímicos**
 - **Aumentar AgroBiodiversidade** (mitigação de pragas e doenças)
- Promover uma gestão integrada dos recursos naturais:
 - Aproveitar as **espécies existentes** (mais robustas e resistentes)
 - **Potenciar a sustentabilidade ambiental e económica**
 - Gerar de novas fontes de rendimento (produtos e negócios)
- Proteger e potencializar as zonas de montanha e flora silvestre
 - Identificar, caracterizar e **monitorizar**
 - Promover a **natureza e biodiversidade**
 - Promover **atividades de lazer e saúde** (turismo, alimentação)
 - Estudar e divulgar as potencialidades das plantas silvestres



Montanha Viva

OBJECTIVO PRINCIPAL:

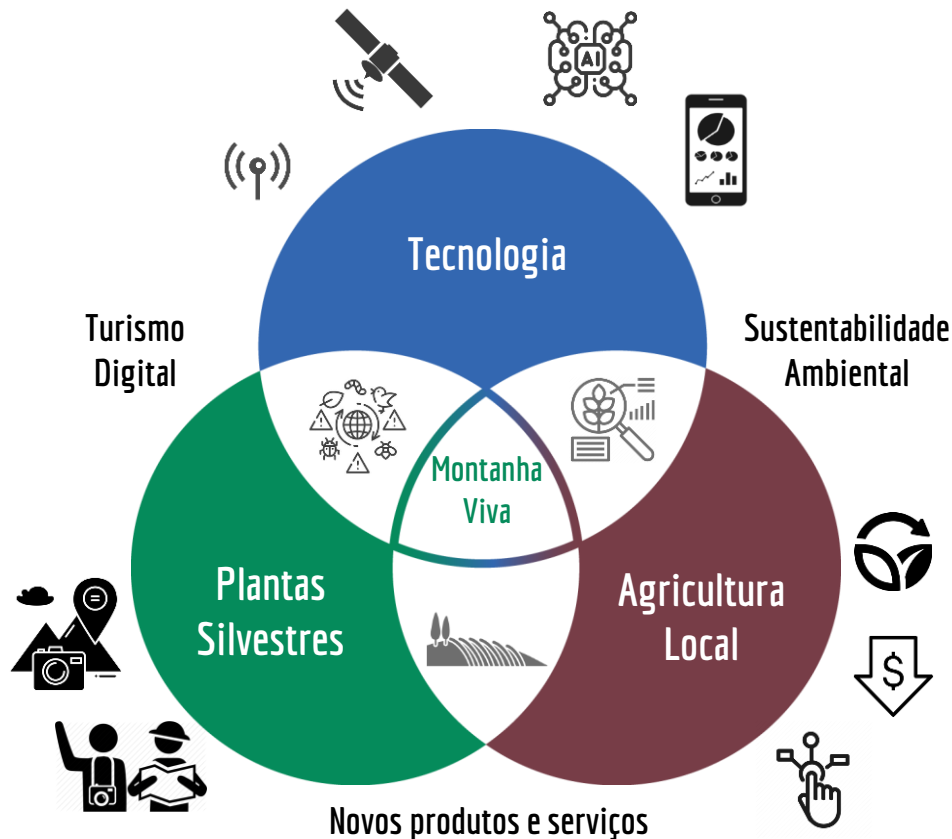
Sistema previsional inteligente de suporte à decisão em sustentabilidade em zonas de montanha que permita potencializar a flora silvestre local.



Montanha Viva

Abordagem Científica e Inovadora:

1. Identificar e estudar as propriedades bioativas das **plantas silvestres** e divulgar os seus usos
2. Implementação de um **sistema de monitorização** inovador e autônomo para zonas remotas.
3. Desenvolvimento de um **sistema de informação inteligente** que permita suportar a agricultura local e incentivar práticas sustentáveis na montanha.
4. **Turismo de Montanha** - Desenvolvimento de percursos pedonais com informações sobre a flora silvestre e seus benefícios.
5. Suporte ao desenvolvimento de **novos negócios**



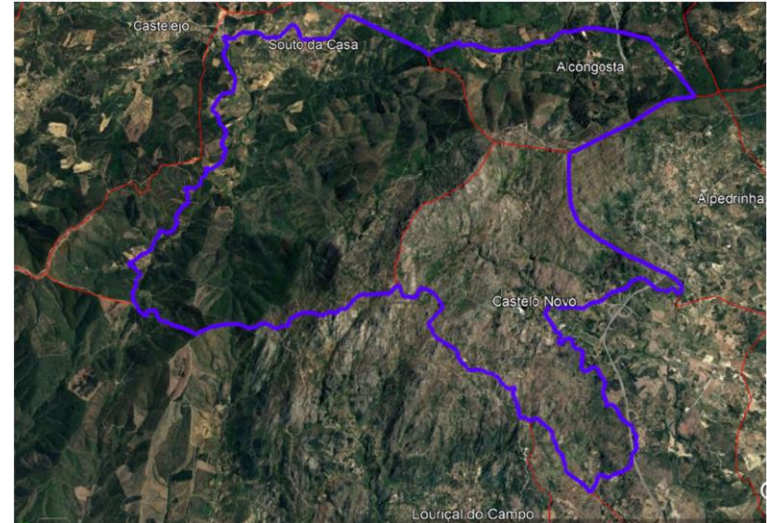
Enquadramento do Projeto

Serra da Gardunha

- Zona Especial de Conservação
- PPRSG (Paisagem Protegida Regional da Serra da Gardunha)
- Distintas floras nas vertentes Norte e Sul

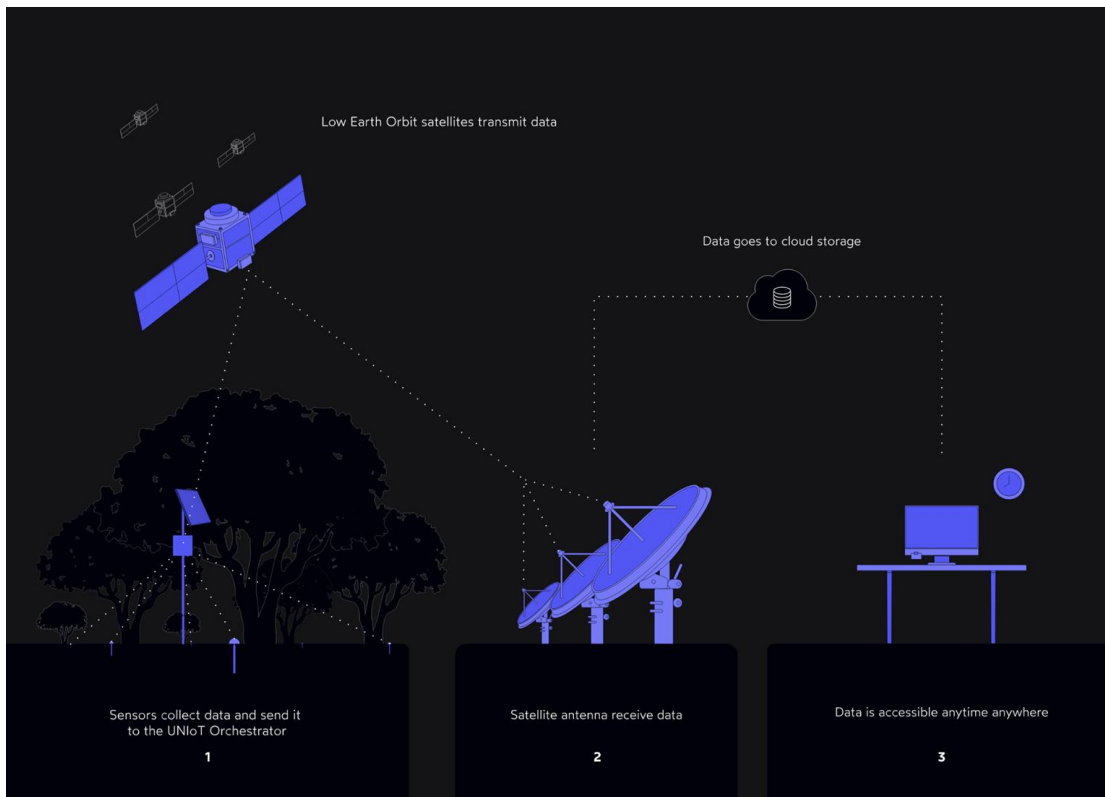
Quintas experimentais:

- QUINTA EXPERIMENTAL DO SEMINÁRIO (20 ha):
Ecossistema de desenvolvimento e validação de tecnologia IoT
- Quintas das Ideias e das Cerejas Ciência Viva (12 ha):
Missão centrada na educação, promoção da cultura científica e valorização dos recursos locais
- Baldios: Testes de novas culturas, soluções tecnológicas e novos negócios.



Tecnologia

1. **Monitorização contínua** das culturas
2. **Comunicação** de dados para o orchestrador
3. **Processamento local** de dados: compressão, encriptação e priorização
4. **Comunicação** entre orchestrador e satélite
5. **Comunicação** entre satélite e gateway
6. Da gateway são distribuídos para a **Cloud**
7. **Processamento, análise e fusão** sensorial na API que integra com a plataforma.
8. **Plataforma de informação** sobre produtividade agrícola, biodiversidade e sustentabilidade ambiental via plataforma.



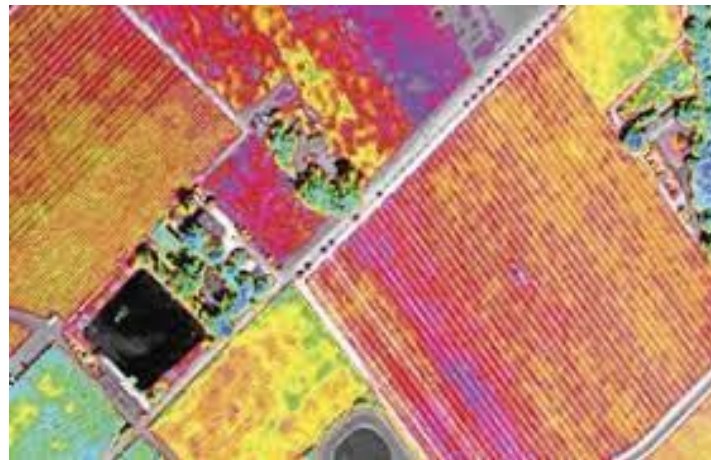
Sensorização Local e GeoEspacial



Ilha de sensores com comunicação IoT



Estação Meteorológica



Sensorização remota por satélite & drones - imagens RGB, multiespectrais e hiperespectrais



Comunicação IoT Via-Satélite



Tecnologia LPWAN (Low Power Wide Area Networking)

Baixo consumo energético

Custos reduzidos

Elevado alcance (satélite)

Elevada cobertura para
terrenos com vasta área

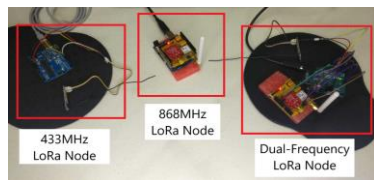
Modular e Adaptável
(sensores e bases de dados)

Capacidade de ligar vários
sensores em simultâneo

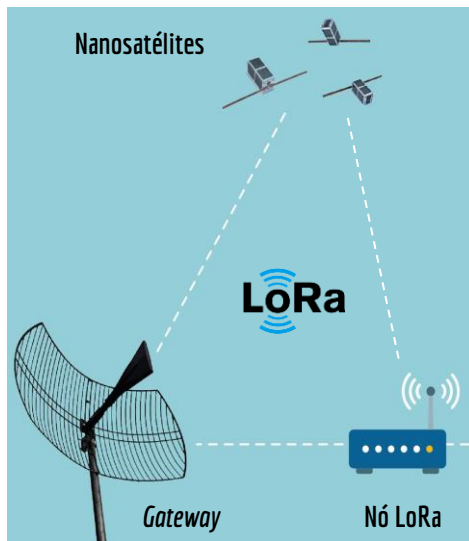
Segurança e Privacidade

Fácil Atualização

Open-source



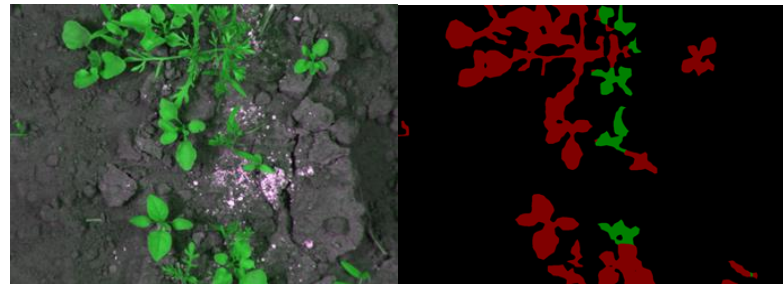
Nós LoRa em desenvolvimento (Spaceway)



Sistema de Informação

Sistema de apoio à decisão baseado em inteligência artificial.

- **Redes Neurais Convolucionais (CNN):** classificação de imagens para deteccção do vigor (crescimento e saúde) das plantas.
- **Redes Neurais de Memória a Longo Curto-Prazo (LSTM):** previsão com base em aprendizagem passada.
- **Priorização Local** da informação a ser enviada por satélite.
- Possibilidade de integração de dados in-situ com dados aéreos, de satélites e históricos (fusão sensorial).



Modelos CNN da UBI para deteccção e classificação de culturas
(vermelho – cultura A; verde – cultura B)

Impactos do Projeto

RECONHECER & MAPEAR

- Estudos etnobotânicos
- Tecnologia de Previsão e comunicação
- Mapear terreno para percursos turísticos



ENVOLVER & CO-CRIAR

- Envolver produtores locais e entidades de proteção ambiental no desenvolvimento da plataforma digital

CAPACITAR & DISSEMINAR

- Plano de comunicação
- Workshops de capacitação para o uso da tecnologia
- Disseminação de boas práticas
- Website, artigos científicos, imprensa, redes sociais

REPLICAR & VALORIZAR

- Raia  e  - similares em flora e ambiente
- Programa de transferência e acompanhamento
- Produto e serviços tecnológicos validados

PARA O TERRITÓRIO E PARA O FUTURO:

- Enriquecimento do património natural (turismo e plantas silvestres)
- Saúde e vitalidade da sociedade
- Ciência e Tecnologia para coesão de um território com elevada desertificação
- Expansão para novas aplicações e utilizadores (e.g outras culturas, floresta, conservação do ambiente, cooperação transfronteiriça)

Sustainable food systems: benefits & opportunities

Healthy and sustainable diets: health and quality of life



Farmers and fishers: fairer prices, sustainable and healthy production practices

New, sustainable business opportunities



Contribute to global transition & future generations

FOOD SECURITY AND SAFETY ARE CORNERSTONES OF OUR FOOD SYSTEM, AND WILL NEVER BE COMPROMISED

#EUFarm2Fork

#EUGreenDeal



2030 Targets for sustainable food production

PESTICIDES



Reduce the overall use and risk of chemical and hazardous pesticides

NUTRIENT LOSSES



Reduce nutrient losses by 50% whilst retaining soil fertility, resulting in 20% less fertilisers

ANTIMICROBIALS



Reduce sales of antimicrobials for farmed animals and aquaculture

ORGANIC FARMING



Increase the percentage of organically farmed land in the EU

#EUFarm2Fork

#EUGreenDeal



Promove
o Futuro
do *Interior*
Concurso 2022



**montanha
viva**

Sistema Previsional Inteligente de Suporte à
Decisão em Sustentabilidade

Gratos pela atenção!



Apoiado pelo Programa Promove da Fundação “la Caixa”, em colaboração com o BPI e com a Fundação para a Ciência e a Tecnologia (FCT)



Fundação
para a Ciência
e a Tecnologia



Fundação “la Caixa”

PHYTOCHEMICAL CHARACTERIZATION AND EVALUATION OF ANTIMICROBIAL PROPERTIES OF WILD PLANTS COLLECTED IN THE MOUNTAIN REGION OF SERRA DA GARDUNHA, PORTUGAL

Alexandra Coimbra¹, Ângelo Luís¹, Pedro Dinis Gaspar², Susana Ferreira¹, Ana Paula Duarte¹

¹ CICS-UBI – Health Sciences Research Centre, University of Beira Interior, Covilhã, Portugal

² C-MAST – Center for Mechanical and Aerospace Science and Technologies, University of Beira Interior, Covilhã, Portugal

Background and Aims: Natural products have received attention as alternative therapeutic options, since they can play a crucial role in the prevention and treatment of different diseases. Therefore, the objective of this work was to evaluate the bioactivity of different plants, studying their chemical composition and bioactivity, namely antimicrobial properties.

Methods: The extracts were obtained with a hydroethanolic solution, with the ethanolic part being evaporated under reduced-pressure and the aqueous part being freeze-dried. Phytochemical characterization was carried out by quantifying total phenolics and flavonoids. The antioxidant activity of the extracts was evaluated by DPPH method and β -carotene/linoleic acid system. The antimicrobial activity of the extracts was evaluated against different Gram-positive and Gram-negative bacteria and yeast.

Results: Six wild plants were identified and collected in the northern area of Serra da Gardunha, *Cistus salvifolius*, *Clinopodium vulgare*, *Glandora prostrata*, *Helichrysum stoechas*, *Rubia peregrina* and *Umbilicus rupestris*. The extracts demonstrated antioxidant activity through different mechanisms. In terms of antimicrobial activity, the extracts mainly exhibited inhibition against Gram-positive bacteria and yeasts, particularly with the *H. stoechas* and *C. salvifolius* extracts demonstrating the strongest activity.

Conclusions: It can be concluded that extracts from these plants demonstrated relevant antioxidant and antimicrobial activities. These results are promising for a possible use of these extracts in the search for new bioactive molecules.

Acknowledgments: Alexandra Coimbra is recipient of a research fellowship (Ref. PD21-00009) within the Research project titled “*Montanha Viva - Sistema Previsional Inteligente de Suporte à Decisão em Sustentabilidade*”, supported by Fundação La Caixa.

Keywords: Plant extracts, phytochemical content, antioxidant activity, antimicrobial activity

Abstract topic:

Antimicrobials and antimicrobial resistance

Phytochemical characterization and evaluation of antimicrobial properties of wild plants collected in the mountain region of Serra da Gardunha, Portugal

Alexandra Coimbra¹, Ângelo Luís¹, Pedro Dinis², Susana Ferreira¹, Ana Paula Duarte¹

¹ CICS-UBI, Health Sciences Research Centre, University of Beira Interior, Covilhã, Portugal

² C-MAST, Center for Mechanical and Aerospace Science and Technologies, University of Beira Interior, Covilhã, Portugal

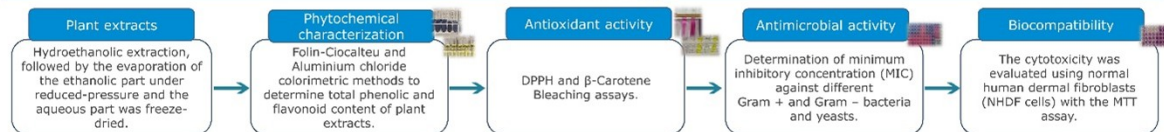
Introduction

Natural products have received attention as alternative therapeutic options, since they can play a crucial role in the prevention and treatment of different diseases. The plants in the mountain region of **Serra da Gardunha** have been used in traditional medicine, attracting interest in their bioactive properties. Seven wild plants were identified and collected in the northern area of Serra da Gardunha, *Cistus salvifolius* aerial parts (**CSAP**) and stems (**CSS**), *Clinopodium vulgare* (**CV**), *Coincya monensis* flowers (**CMF**) and stems (**CMS**), *Glandora prostrata* (**GP**), *Helichrysum stoechas* (**HS**), *Rubia peregrina* (**RP**) and *Umbilicus rupestris* flowers (**URF**) and leaves (**URL**).

Aim: This work aimed to evaluate the bioactivity of different plants, studying their chemical composition and bioactivity, namely antimicrobial properties.



Methodology



Results

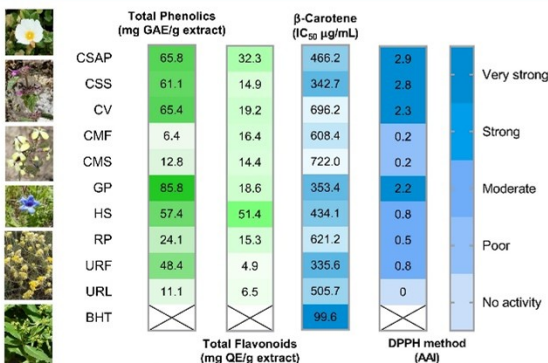


Figure 1. Phytochemical characterization and Antioxidant activity evaluation. AAI - Antioxidant activity index; BHT - Butylated hydroxytoluene, used as positive control; DPPH - 2,2-diphenyl-1-picrylhydrazyl; GAE - gallic acid equivalents; QE - quercetin equivalents.

The GP extract contains a higher total phenolic content and the HS extract has a greater total flavonoids content.

The extracts have antioxidant activity through different mechanisms. The DPPH methodology and based on Scherer and Godoy (2009) classification, showed that the CSAP, CSS, CV and GP extracts exhibited very strong antioxidant activity related to scavenge free radicals. The β-carotene bleaching assay demonstrated that the CSS, GP and URF extracts can inhibit lipid peroxidation.



Figure 2. Biocompatibility for NHDf cell line measured by MTT assay after 24 h of treatment with the extracts. The results are presented as IC₅₀ values in mg/mL.

The viability of NHDf cells exceeded 75% when exposed to extracts at concentrations of 1 mg/mL or lower, except for the CSAP and HS extracts (IC₅₀ < 1 mg/mL).

Table 1. Antimicrobial activity evaluated by the determination of the MIC values (mg/mL) in bacterial and yeast species (modal values).

Species	CSAP	CSS	CV	GP	HS	URF	URL
<i>Staphylococcus aureus</i> ATCC 25923	0.5	0.5	>2	2	0.008	>2	>2
<i>Staphylococcus aureus</i> MRSA 05/15	0.5	0.5	>2	>2	0.008	>2	>2
<i>Bacillus cereus</i> ATCC 11778	0.25	0.5	>2	>2	0.008	>2	>2
<i>Listeria monocytogenes</i> LMG 16779	0.5	1	>2	>2	0.008	>2	>2
<i>Escherichia coli</i> ATCC 25922	>2	>2	>2	>2	>2	>2	>2
<i>Klebsiella pneumoniae</i> ATCC 13883	1	1	2	1	2	>2	2
<i>Pseudomonas aeruginosa</i> ATCC 27853	>2	>2	>2	>2	>2	>2	>2
<i>Salmonella</i> Typhimurium ATCC 13311	>2	>2	>2	>2	>2	>2	>2
<i>Acinetobacter baumannii</i> LMG 1025	2	2	>2	>2	2	>2	>2
<i>Acinetobacter baumannii</i> AcB 13/10	1	1	>2	>2	2	>2	>2
<i>Candida albicans</i> ATCC 90028	0.03	0.02	>2	>2	0.5	0.03	>2
<i>Candida tropicalis</i> ATCC 750	0.25	0.25	1	1	0.5	0.25	>2

The antimicrobial evaluation demonstrated that the HS extract exhibited the strongest antibacterial activity against the Gram-positive, while the CSAP and CSS extracts also showed activity, albeit at comparatively lower values. Additionally, the CSAP, CSS and URF extracts showed a promising anti-*Candida* activity.

Conclusion

It can be concluded that extracts from these plants demonstrated relevant bioactive activities. These results are promising, indicating the potential of these extracts in the discovery of new bioactive molecules.

Acknowledgements

Research project financed by "Montanha Viva - Sistema Previsional Inteligente de Suporte à Decisão em Sustentabilidade" (Ref. PD21-00009), supported by Fundação La Caixa, through the program "Promove o Futuro do Interior, Concurso 2022".



Look at the project
"Montanha Viva"





TECHNOLOGICAL INNOVATIONS FOR REMOTE MONITORING AND AI-BASED DECISION SUPPORT IN MOUNTAIN ECOSYSTEMS: THE MONTANHAVIVA PROJECT

GASPAR, P.D. (1); AGUIAR, M. (1); PEREIRA, N. (1); ANTUNES, R. (1); SOUSA, M. (1); VELOSO, M.; FERREIRA, D. (1); ALVES, D. (1); ALVES, A.C. (1); CORCEIRO, A. (1)

(1) C-MAST - Centre for Mechanical and Aerospace Science and Technologies, University of Beira Interior, Portugal E-mail: dinis@ubi.pt; martim.aguiar@ubi.pt; nuno.pereira@ubi.pt; rodrigo.antunes@ubi.pt; galvao.sousa@ubi.pt; mariana.veloso@ubi.pt; daniel.b.ferreira@ubi.pt; david.filipe.alves@ubi.pt; ana.cristina.alves@ubi.pt; ana.corceiro@ubi.pt

ENDEREÇO DE CORRESPONDÊNCIA: dinis@ubi.pt

RESUMO

O projeto MontanhaViva integra tecnologias avançadas para monitorizar e apoiar o turismo sustentável em áreas montanhosas. Os principais desenvolvimentos incluem sistemas de deteção remota, análise de imagens de alta resolução e uma plataforma inteligente para prever o crescimento das plantas. O projeto utiliza comunicação sem fios para transferência de dados, redes de sensores para monitorização em tempo real e algoritmos de IA para avaliar o vigor das plantas. Uma plataforma web consolida estas tecnologias, permitindo a visualização de dados e apoio à decisão. Adicionalmente, são propostas oportunidades para o desenvolvimento de novos negócios centrados no turismo sustentável e na comercialização de produtos regionais. Uma inovação crucial é a aplicação móvel que guia os turistas por trilhos de montanha com flora endémica, fornecendo informações em tempo real sobre as propriedades medicinais e aromáticas das flores nativas, melhorando a experiência do visitante. Esta tecnologia apoia a criação de visitas guiadas, linhas de produtos ecológicos e workshops educacionais. Ao integrar IA e serviços de geolocalização, a aplicação promove uma experiência de turismo sustentável e enriquecedora em ecossistemas de montanha. Esta abordagem tecnológica oferece insights valiosos para a gestão de ecossistemas, alinhando-se com os objetivos de turismo sustentável.

Palavras-chave: Deteção Remota, Suporte à Decisão com IA, Turismo Sustentável, Monitorização de Ecossistemas, Conservação da Biodiversidade.

ABSTRACT

The MontanhaViva project integrates advanced technologies to monitor and support sustainable mountain tourism. Key developments include remote sensing systems, high-resolution image analysis, and an intelligent platform for predicting plant growth. The project employs wireless communication for data transfer, sensor networks for real-time monitoring, and AI algorithms to assess plant vigor. A web-based platform consolidates these technologies, enabling data visualization and decision-making support. Additionally, opportunities for developing new businesses focused on sustainable tourism and the commercialization of regional products are proposed. A key innovation is the mobile application that guides tourists through mountain paths featuring endemic flora. The app provides real-time information on the medicinal and aromatic properties of native flowers, enhancing the visitor experience. This technology supports the creation of guided tours, eco-friendly product lines, and educational workshops. By integrating AI and geolocation services, the app fosters an enriched and sustainable tourism experience in mountain ecosystems. This technological approach offers valuable insights into ecosystem management, aligning with sustainable tourism goals.

Keywords: Remote Sensing, AI-based Decision Support, Sustainable Tourism, Ecosystem Monitoring, Biodiversity Conservation.



IMPLEMENTATION OF INTERACTIVE DIGITAL PANELS TO PROMOTE ENVIRONMENTAL SUSTAINABILITY AND TOURISM IN SERRA DA GARDUNHA: THE MONTANHA VIVA APPROACH

FERREIRA, D. (1); SOUSA, M. (1); CORCEIRO, A. (1); VELOSO, M.; ALVES, D.
(1); ALVES, A.C. (1); AGUIAR, M. (1); ANTUNES, R. (1); PEREIRA, N. (1);
GASPAR, P.D. (1)

(1) C-MAST - Centre for Mechanical and Aerospace Science and
Technologies, University of Beira Interior, Portugal E-mail:
daniel.b.ferreira@ubi.pt; galvao.sousa@ubi.pt; ana.corceiro@ubi.pt;
mariana.veloso@ubi.pt; david.filipe.alves@ubi.pt;
ana.cristina.alves@ubi.pt; martim.aguiar@ubi.pt; rodrigo.antunes@ubi.pt;
nuno.pereira@ubi.pt; dinis@ubi.pt

ENDEREÇO DE CORRESPONDÊNCIA: dinis@ubi.pt

RESUMO

A região da Serra da Gardunha possui uma rica biodiversidade e um potencial significativo para a educação ambiental e o turismo sustentável. O projeto Montanha Viva pretende explorar este potencial, melhorando as experiências dos visitantes ao longo dos trilhos pedestres aí localizados. Este estudo visa desenvolver e implementar painéis digitais interativos ao longo dos percursos pedestres dos trilhos pedestres. Os painéis são concebidos para fornecer aos visitantes informações detalhadas sobre a flora local, expondo propriedades bioativas e potenciais aplicações na saúde, promovendo assim a sustentabilidade ambiental e o turismo.

O processo de desenvolvimento compreendeu duas etapas. Na primeira etapa, foi desenhada uma estrutura externa e os componentes foram fabricados utilizando tecnologia de corte a laser, sendo posteriormente montados numa unidade coesa baseada num design inovador e moderno. Na segunda etapa, foi integrada uma interface digital, com um monitor para apresentar o conteúdo interativo. Foi desenvolvida uma apresentação de fácil utilização, permitindo aos visitantes aceder a informações sobre a flora circundante e percursos pedestres. Cinco botões táteis foram incorporados na estrutura externa do painel para facilitar a navegação intuitiva e o acesso rápido à informação desejada.

Os painéis digitais interativos conseguiram fundir com sucesso o design estético com a utilidade funcional. Os visitantes podem agora aceder a informações abrangentes sobre a flora local, incluindo as suas propriedades bioativas e usos relacionados com a saúde e bem-estar,

aumentando tanto o valor educativo como o envolvimento dos visitantes.

Espera-se que a implementação destes painéis interativos melhore significativamente a educação ambiental e a experiência dos visitantes ao longo dos trilhos pedestres da Serra da Gardunha. Esta iniciativa enquadra-se nos objetivos do projeto Montanha Viva de promover a sustentabilidade ambiental e melhorar o turismo na região da Serra da Gardunha.

Palavras-chave: Turismo sustentável, Painéis interativos, Sustentabilidade ambiental, Envolvimento do visitante.

ABSTRACT

The Serra da Gardunha region has rich biodiversity and significant potential for environmental education and sustainable tourism. The Montanha Viva project aims to explore this potential by enhancing visitor experiences along the Gardunha trails. This study aims to develop and implement interactive digital panels along the walking routes of the Gardunha trails. The panels are designed to provide visitors with detailed information about the local flora, emphasizing bioactive properties and potential health applications, thereby promoting environmental sustainability and tourism.

The development process comprised two stages. In the first stage, an external structure was designed and components were crafted utilizing laser-cutting technology and assembled into a cohesive unit based on an innovative and modern design. In the second stage, a digital interface was integrated, featuring a monitor to display interactive content. A user-friendly presentation was developed, allowing visitors to access information on surrounding flora and hiking routes. Five tactile buttons were incorporated into the panel's outer structure to facilitate intuitive navigation and quick access to desired information.

The interactive digital panels successfully merged aesthetic design with functional utility. Visitors can now access comprehensive information about local flora, including their bioactive properties and health-related uses, enhancing both educational value and visitor engagement.

The implementation of these interactive panels is expected to significantly enhance environmental education and visitor experience along the Gardunha trails. This initiative is within the Montanha Viva project's objectives of promoting environmental sustainability and improving tourism in the Serra da Gardunha region.

Keywords: Sustainable tourism, Interactive panels, Environmental sustainability, Visitor engagement.

Evaluation of the bioactive activities of wild plants from Serra da Gardunha mountain region in Portugal

Alexandra Coimbra^{1,2}, Ângelo Luis^{1,2}, Pedro Dinis³, Susana Ferreira^{1,2}, Ana Paula Duarte^{1,2}

¹ CICS-UIB – Health Sciences Research Centre, University of Beira Interior, Covilhã, Portugal

² RISE-Health – Department of Medical Sciences, Faculty of Health Sciences, University of Beira Interior, Av. Infante D. Henrique, 6200-506 Covilhã, Portugal

³ C-MAST – Center for Mechanical and Aerospace Science and Technologies, University of Beira Interior, Covilhã, Portugal

INTRODUCTION

Natural products have received attention as alternative therapeutic options, since they can play a crucial role in the prevention and treatment of different diseases. The **Montanha Viva** project proposes to collect information about the Serra da Gardunha region's wild plants used by local communities and to investigate potential new phytotherapeutic compounds with antioxidant activity and/or for the treatment of microbial-related pathologies. Two wild plants were identified and collected in the northern area of Serra da Gardunha as presenting potential bioactive properties, *Cistus salvifolius* aerial parts (**CSAP**) and stems (**CSS**) and *Helichrysum stoechas* (**HS**).

AIM: Therefore, the aim of this study was to assess the bioactivity of these plants, with a focus on antimicrobial activity.

METHODOLOGY



Plant extracts

Hydroethanolic extraction, followed by the evaporation of the ethanolic part under reduced-pressure and the aqueous part was freeze-dried.

Antioxidant activity

DPPH and β -Carotene Bleaching assays.



Anti-inflammatory activity

Inhibition of protein denaturation using bovine serum albumin (BSA)

Antimicrobial activity

Determination of minimum inhibitory concentration (MIC) against different Gram + and Gram – bacteria and yeasts.

Antiviral activity

Inhibition of biofilm formation through crystal-violet assay.



Combined effect with antibiotics

Checkerboard Assay



RESULTS

Table 1. Antioxidant activity using the DPPH method and β -carotene-bleaching assay and **anti-inflammatory activity** results (results expressed as mean \pm standard deviation).

Samples	DPPH method			β -Carotene-Bleaching Assay	Anti-Inflammatory Activity
	IC ₅₀ (μ g/mL)	AAI	Classification	IC ₅₀ (μ g/mL)	IC ₅₀ (μ g/mL)
CSAP	19.18 \pm 5.43	2.93 \pm 0.12	Very strong	466.23	745.34 \pm 31.96
CSS	20.72 \pm 5.73	2.84 \pm 0.28	Very strong	342.68	700.42 \pm 14.87
HS	67.83 \pm 18.87	0.83 \pm 0.04	Moderate	434.12	51.75 \pm 3.76
Gallic acid	3.92 \pm 1.26	13.00 \pm 0.67	Very strong	-	-
BHT	-	-	-	99.56	-
Acetylsalicylic acid	-	-	-	-	4.20 \pm 1.41

AAI – Antioxidant activity index; BHT – Butylated hydroxytoluene; DPPH – 2,2-diphenyl-1-picrylhydrazyl.

The **DPPH methodology** showed that the CSAP and CSS extracts exhibited very strong antioxidant activity related to scavenge free radicals.

The **β -carotene bleaching assay** demonstrated that the CSS extract can inhibit lipid peroxidation.

Table 2. Antimicrobial activity evaluated by the determination of the MIC values (mg/mL) in bacterial and yeast species (modal values).

Species	CSAP	CSS	HS
<i>Staphylococcus aureus</i> ATCC 25923	0.5	0.5	0.008
<i>Staphylococcus aureus</i> MRSA 12/08	> 2	2	0.06
<i>Bacillus cereus</i> ATCC 11778	0.25	0.5	0.008
<i>Listeria monocytogenes</i> LMG 16779	0.5	1	0.008
<i>Klebsiella pneumoniae</i> ATCC 13883	1	1	2
<i>Acinetobacter baumannii</i> LMG 1025	2	2	2
<i>Acinetobacter baumannii</i> AcB 13/10	1	1	2
<i>Candida albicans</i> ATCC 90028	0.03	0.02	0.5
<i>Candida tropicalis</i> ATCC 750	0.25	0.25	0.5

The extracts exhibited **antimicrobial activity**, particularly against *S. aureus* and *Candida* species presenting MIC values between 0.008 and 0.5 mg/mL.

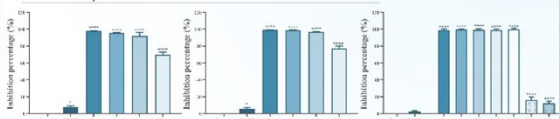


Figure 1. Effect of different concentrations of the extracts on the **formation of biofilms** of *S. aureus* ATCC 25923 strain and results are expressed as % of biofilm biomass inhibition.

CONCLUSION

These findings are encouraging for the potential use of these extracts in the discovery of new bioactive compounds and the possible use of the extracts as an alternative antibacterial agent for the control of *Staphylococcus aureus*.



Look at the project:



ACKNOWLEDGMENTS

Alexandra Coimbra is recipient of a research fellowship (Ref. PD21-00009) within the Research project titled "Montanha Viva - Sistema Previsional Inteligente de Suporte à Decisão em Sustentabilidade, supported by Fundação La Caixa. This work was developed within the scope of the CICS-UIB projects UIDB/00709/2020 and UIDP/00709/2020, financed by national funds through the Portuguese Foundation for Science and Technology/MCTES.

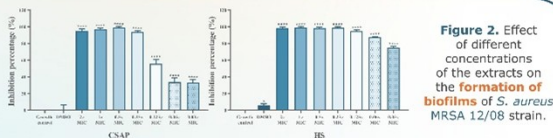


Figure 2. Effect of different concentrations of the extracts on the **formation of biofilms** of *S. aureus* MRSA 12/08 strain.

The extracts demonstrated **anti-viral activity**, inhibiting the formation of biofilms of *S. aureus* strains even at subinhibitory concentrations.

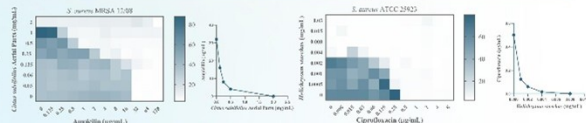


Figure 3. Some examples of the results obtained for the **combined effect with antibiotics**. **Checkerboards** of CSAP and ampicillin (FICI=0.19–0.28) or HS and ciprofloxacin (FICI=0.37–0.5) for growth inhibition of *S. aureus* MRSA 12/08 and *S. aureus* ATCC 25923, respectively. In the checkerboard graphics, white indicates 0% cell viability using the resazurin assay. Points on **isobolograms** represent combinations of extracts and antibiotics (relative to their MICs alone).

Table 3. Effect of the interaction between extracts and antibiotics based on the fractional inhibitory concentration index (FICI) value calculated from the checkerboard assay.

Antibiotics	<i>S. aureus</i> strains	CSAP	CSS	HS
Ampicillin	ATCC 25923	Additive	Additive	Additive
	SA 03/10	Synergistic	Synergistic	No interaction
	MRSA 12/08	Synergistic	Synergistic	Synergistic
Ciprofloxacin	ATCC 25923	Additive	Synergistic	Synergistic
	SA 03/10	No interaction	No interaction	Additive
	MRSA 12/08	Synergistic	Synergistic	Synergistic
Vancomycin	ATCC 25923	Additive	No interaction	Additive
	SA 03/10	-	-	-
	MRSA 12/08	Synergistic	Additive	Additive

"-" was used when we obtained growth across the entire plate not allowing to calculate the FICI.

The combination of the extracts with the antibiotic's ampicillin, ciprofloxacin or vancomycin **potentiated the effects of these antibiotics** against the *S. aureus* strains tested, demonstrating mostly synergistic (FICI \leq 0.5) or additive (0.5 < FICI \leq 1) effect. Some of these interactions led to a resensitization of the resistant strains to these antibiotics.