Computers and Electronics in Agriculture 218 (2024) 108712

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Biomass characterization with semantic segmentation models and point cloud analysis for precision viticulture

A. Bono [a], R. Marani [b,*], C. Guaragnella [a], T. D'Orazio [b]

[a] Polytechnic University of Bari, Department of Electrical and Information Engineering (DEI), Via E. Orabona 4, Bari, Italy
[b] Institute of Intelligent Industrial Systems and Technologies for Advanced Manufacturing (STIIMA), National Research Council (CNR), Via Amendola 122, D/O, Bari, Italy

ARTICLE INFO

Keywords:
Precision viticulture
Image semantic segmentation
Point cloud processing
Removed leaf detection

ABSTRACT

The scientific progress in artificial intelligence and robotics has enabled precision viticulture to pursue sustainability and improve the final yield. For instance, monitoring the canopy volume of each plant can allow the correct ripening of the bunches. In this context, this paper proposes a novel approach for the characterization of biomass volume using images acquired in a vineyard with the low-cost Azure Kinect RGB-D camera. Semantic image segmentation is implemented using three encoder–decoder deep architectures (U-Net, DeepLabV3+, and MANet) to produce accurate masks of the vine leaf structure. In a transfer learning approach, a public dataset acquired with the Intel RealSense D435 depth camera is used to train the segmentation networks. Then, a complete pipeline to estimate possible changes in biomass volume is presented. Experiments are run to analyze the biomass removed during the trimming process of grapevine plants. The best segmentation result is obtained by the U-Net architecture with ResNet50 backbone, showing an accuracy of 92.10%, although the training and test sets consist of images acquired by different cameras. However, the DeepLabV3+ network with ResNeXt50 backbone, which scores an accuracy of 90.25% on the test set, gives the best estimate of the removed biomass, requiring the shortest time for training. These outcomes prove the potential capability of this automatic approach for controlling leaf growth and ensuring sustainable viticulture practices.

1. Introduction

In this century, sustainability, food security, smart use of agricultural resources, crop yield, and quality improvement are the basis of new agricultural paradigms, which account for the spatial and temporal variability of crop and soil characteristics within actual fields (Stafford, 2000). Precision viticulture (PV) is a subset of precision agriculture (PA) and aligns with its core objectives (Arnó Satorra et al., 2009). Both aim to manage crops effectively, enhance economic benefits, and minimize environmental impact (Matese et al., 2015).

Viticultural areas are characterized by an irregular spatial distribution that involves complex maintenance and control activities for farmers. Automation applied to precision farming allows different cultivation practices within different areas of the vineyard (Comba et al., 2018; King et al., 2014; Bramley and Hamilton, 2004). Among these practices, trimming is fundamental during the phenological phases. It regulates vegetation volume to maintain microclimatic conditions inside the canopy, such as penetration of light and air circulation, to disfavor the development of diseases and allow the correct ripeness of the bunches. Because of the importance of this operation, in this paper, we propose an approach for estimating the plant volume changes after the trimming process.

In this context, the development of Artificial Intelligence (AI), smart sensors, and robotics offers non-invasive approaches to assessing several plant characteristics, such as canopy volume, plant height, leaf area coverage, and biomass (Botta et al., 2022). Remote Sensing (RS) (Cisternas et al., 2020) offers rapid and comprehensive information about the morphology, dimensions, and vitality of grapevines across entire vineyards (Hall et al., 2002; Oliver et al., 2013). The fundamental elements of remote sensing are platforms and sensors. The leading platforms are satellites, aerial (aircraft and unmanned aerial vehicles, UAVs), and ground-based platforms (unmanned ground vehicles, UGVs) (Jafarbiglu and Pourreza, 2022). The choice of a platform depends on the specific application, as each of them has advantages and disadvantages.

Satellite platforms have several limitations due to their dependence on meteorological conditions, high costs, and the difficulty in discerning the composition of vineyards characterized by inter-row paths and herbaceous vegetation (Borgogno-Mondino et al., 2018; Sishodia et al., 2020; Matese and Filippo Di Gennaro, 2015). Low-altitude platforms,

* Corresponding author.
E-mail address: roberto.marani@stiima.cnr.it (R. Marani).

https://doi.org/10.1016/j.compag.2024.108712
Received 15 June 2023; Received in revised form 19 December 2023; Accepted 3 February 2024
Available online 12 February 2024

such as manned or unmanned aerial vehicles, provide high-resolution images, making it easier to distinguish vines from weeds and other objects. Manned aircraft guarantees a better spatial resolution, an arbitrary frequency of observations, and real-time availability of raw data, but are costly and inflexible in flight scheduling due to airspace regulations (Matese et al., 2015). UAVs, though cheaper, can cover smaller areas, but offer high-resolution imagery, helping to differentiate canopy pixels and classify details within canopies (Jafarbiglu and Pourreza, 2022).

The last type of platform is the ground-based one, which is also called proximal due to the closer position of sensors, within a few meters, to the target (e.g., soil, plant, crop, etc.) than in the other platforms (Sishodia et al., 2020). The advantages over space and airborne remote sensing are portability, flexibility, and controllability. These platforms can be mobile, if they carry out measurements while moving or "on-the-go" being installed directly on machines, such as tractors or agriculture robots, or they can be fixed if the measurements are taken in a fixed position, e.g. by using tripods. Proximal sensing is suitable for both small and large-scale monitoring, offering high-resolution images without constraints like flight schedules and climatic conditions (Andújar et al., 2019; Moreno and Andújar, 2023).

Each of the mentioned platforms can be equipped with sensors of different types; typical sensors are optoelectronic, such as LiDAR and RGB-D. (Pallottino et al., 2019; Vulpi et al., 2022). This type of sensor can be passive or active about the energy source (Oliver et al., 2013). Using a laser beam, LiDARs allow mapping the field to perform tasks such as phenotyping (Lin, 2015); they also provide 3D point cloud, but differences in the color and shapes of the canopy and ambient light changes can produce outliers in their 3D models. RGB-D cameras have been proposed as an alternative solution to recover 3D colored models of plants. Depth cameras can operate on Time-of-Flight (ToF) or Stereo Vision (SV) principles. The ToF cameras are more precise in case of challenging lighting conditions than SV sensors, which are better in terms of 3D images (Moreno and Andújar, 2023). Due to their low cost and ease of use, the diffusion of RGB-D cameras allowed digital imaging and computer vision development using Machine Learning (ML) techniques, making proximal sensing powerful in characterizing vine traits. The work in Mohimont et al. (2022) discusses the application of ML techniques in viticulture which mainly focus on the detection, counting and prediction of grape yield. Several methods in image analysis use Convolutional Neural Networks (CNN) for image processing that consist of developing segmentation, shape recognition, and feature extraction algorithms starting from natural images. For example, in Palacios et al. (2020), an algorithm for grapevine flower counting is developed to forecast crop yield. The prompt detection of diseases can also increase the sustainability of crops, as the use of pesticides can be dramatically reduced. In Kerkech et al. (2020), an approach to segmenting UAV images is proposed to map diseased areas and guarantee the healthy state of the plant by continuous monitoring. Plant phenotyping is another critical aspect. In Milella et al. (2019), RGB-D cameras are used to produce data, which CNNs process for automating grapevine phenotyping. In this way, canopy volume estimation and bunch counting can be effectively approached. Further developments in Marani et al. (2021) propose an automated segmentation of grape bunches in color images acquired from an RGB-D camera mounted on an agricultural vehicle. One of the main bottlenecks of deep learning is the need for large sets of labeled data and powerful computing resources. The recent availability of public datasets, manually labeled (Santos et al., 2020; Apostolidis et al., 2022; Casado-García et al., 2023), can enable their use, even for biomass characterization, which plays a central role in vineyard management. A discussion point concerns the question that in literature individual datasets are often relied upon for both the training and testing phases, especially in the agricultural sector where datasets focus on specific phenological phases of plants. In this context, it can be very useful to develop a model capable of generalization, as proposed in Casado-García et al. (2023), in which an approach for bunch detection was proposed involving the use of different datasets, cameras and grape varieties.

Fig. 1. An agricultural machine trimming the grapevines in the Conte Spagnoletti Zeuli vineyard (Italy) where the field experiments were done.

However, to the best of the authors' knowledge, depth data captured on the field and deep learning techniques have not been used for a complete and detailed study of the vine plant canopy and the estimation of volume changes in time. In fact, for example in Di Gennaro and Matese (2020) the volume of vine biomass was evaluated through high-resolution images acquired by UAV and point cloud analysis without the application of DL techniques. On the contrary, in other contexts, the study of the biomass volume was done. In Liu et al. (2021) RGB and depth images were used to estimate the canopy growth of the Toona sinensis plant. Another example is Qi et al. (2021) where deep learning networks and point cloud models from UAV imagery are used to segment and calculate the canopy volume of Citrus reticulate Blanco cv. Shatangju trees, determining optimal methods for accurate volume estimation. Although there are works on leaf volume combining deep learning techniques and 3D imaging, a lack of studies emerges for the characterization of the vine's biomass also considering that the acquisition part is often carried out using UAV platforms.

This work proposes a comprehensive approach for estimating the volume of leaves removed after the trimming process using color and depth information provided by an RGB-D camera to help farmers monitor plant condition and the resulting expected production amount. In detail, for the control of this process, the use of a low-cost system based on a ground-based platform has been proposed, represented by the RGB-D camera Azure Kinect product by Microsoft, which could be integrated into a tractor that normally performs vineyard maintenance operations. Semantic segmentation techniques and point cloud analysis were used to extract the desired information.

The main proposed innovations are the following:

1. The implementation of semantic segmentation was done by exploiting two different datasets for the training and testing phases, respectively. The two datasets were acquired by different cameras, namely the Intel RealSense D435 and the Microsoft Azure Kinect, mounted on different setups, framing the "Negroamaro" and the "Nero di Troia" Vitis Vinifera cultivars. The dataset acquired with the Microsoft Azure Kinect, considered for the biomass volume evaluation after the trimming process, consists of very limited data (36 images) that are insufficient for training any methodologies but are enough for testing purposes. Therefore, the transfer learning approach aims to overcome the problem of scarcity of data, find a solution to the demand for a large amount of labeled data needed for training neural

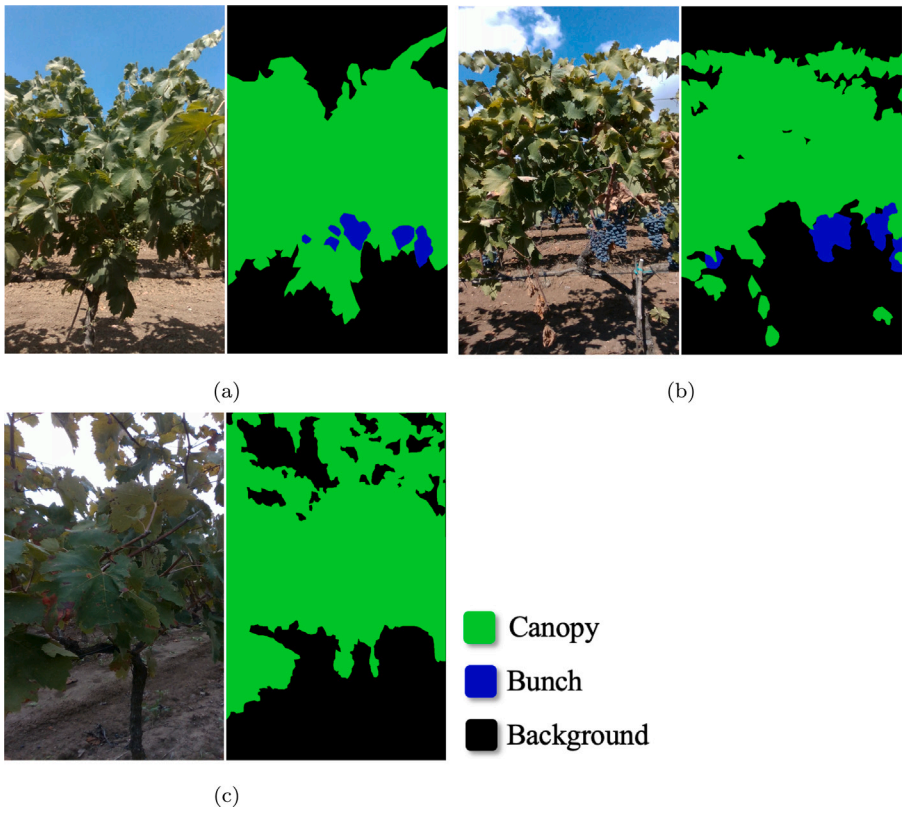

Fig. 2. Color images and corresponding ground truth mask acquired with the Intel RealSense D435 camera (a) in July, when the grapes were small and green and there were few leaves, (b) in September, when the grapes were red and there were many leaves, and (c) in October, when there were few leaves and no grapes. The legend shows the three classes of interest.

networks, and try to develop a model that can generalize and be applied in as many contexts as possible. In addition, the dataset used in this paper is innovative compared to existing public datasets, as it allows the exploitation of all the sensor's multimodal data, a crucial factor for the present work. This approach enables the comprehensive analysis of volumetric changes in leaf mass, a dimension that remains unexplored when relying only on RGB images typically found in prevalent public datasets. As highlighted in Blekos et al. (2023), a substantial gap exists in viticulture-related public datasets, wherein, as demonstrated by Lu and Young (2020), the majority comprise RGB and labeled images.

2. The multimodality of the sensor was exploited to correlate the information extracted from the semantic segmentation with the estimate of the volume information provided by the 3D point clouds obtained from natural images in a real context. In particular, by comparing the 3D images corresponding to the segmented area extracted with the deep learning techniques, it was possible to have the value of the weight of the leaves removed from the trimming process and to compare it with the real weight.
3. A registration procedure was used to superimpose the point clouds and allow volume comparisons. In this way, since the proposed approach has no constraints on the sensor's pose with respect to the plants, the proposed pipeline can allow a future camera integration into standard tractors performing maintenance operations.

Interesting results were obtained both from the semantic segmentation and from the analysis of the point clouds, thus highlighting how combining these two techniques allows the automatic estimation of the removed biomass and the control of the volume of the vine canopy to guarantee correct ripening of the bunches.

The paper is organized as follows: Section 2 describes the datasets, the deep neural networks for image semantic segmentation and the techniques for point cloud processing; Section 3 presents the experiments and the results of the processing; Section 4 proposes a discussion of the results obtained in the study by addressing its limitations in order to be able to offer explanations on the observed results and future improvements; Section 5 reports final comments.

2. Materials and methods

The framework of this paper consists of combining image semantic segmentation and point cloud analysis to estimate the leaf volume removed during trimming (Fig. 1) using images acquired with the Microsoft Azure Kinect camera. The next subsections detail the whole framework, also presenting the on-field datasets.

2.1. Datasets

In this work, to detect the canopy of grapevine plants a semantic segmentation approach was applied by using two different datasets: RGB images of a public dataset acquired by the Intel RealSense D435 were used for training the models; RGB images acquired by the Microsoft Azure Kinect during on-field experiments were considered for the testing phase. This transfer learning approach was partially based on the work of Casado-García et al. (2023), where the experiments were run by training and testing the models with the same dataset and by evaluating the generalization capacity on different datasets. In the proposed approach, the innovation is to exploit two completely different datasets for the training and testing phases.

The following subsections describe these datasets, highlighting their differences.

2.1.1. Training dataset

A public dataset[1] obtained with the Intel RealSense D435 depth camera was used as the training set. In general, this camera provides color, infrared and depth images; but in this case, only color ones were used during training.

The images were acquired in a vineyard in San Donaci (Italy) where the Vitis Vinifera, cultivar "Negroamaro", is grown as a grape variety. The camera was tilted by 90° to have the data in portrait orientation and was mounted on a moving robot that caught lateral views of the row at a distance of 0.8 to 1 m covering a horizontal FoV between 0.9 and 1.2 m. In this way, every plant is framed in a single picture. The dataset consists of 315 color images in PNG format with 1280 × 720-pixel resolution, as the ones shown in Fig. 2, captured during different phenological phases of the vineyard. This dataset had pre-labeled images that create a segmentation ground truth, shown in Fig. 2, of three classes: canopy, bunches and background. The same dataset was used in Casado-García et al. (2023) by selecting a larger part for training and the remaining part for testing the networks. This work aimed to demonstrate that a model constructed on only a type of data could be used to evaluate different contexts. For this reason, the whole dataset acquired with the RealSense was used during the training phase, while for the testing a different dataset was considered. In detail, the RealSense D435 dataset was randomly divided into the training and validation sets, considering that 20% of the data was reserved for validation and the remaining 80% was used for training. A specific function, i.e. the Shuffle function (W3schools, 2023), was employed to mix these two portions of the dataset at each epoch, preserving the initial division while altering the data distribution within both subsets. This approach aimed to prevent the model from learning the order of the data and ultimately enhancing its ability to generalize.

2.1.2. Testing dataset

The testing dataset[2] consisted of 36 images that were used to verify the ability of the network to generalize and produce a good segmentation even on different images from those used in the training phase. The images were acquired on June 23rd 2022 in the Conte Spagnoletti Zeuli vineyard in Andria (Italy), in which the Vitis Vinifera, cultivar "Nero di Troia" (red wine grape variety) is grown. The Microsoft Azure Kinect camera was mounted horizontally on a tripod fixed at a variable distance between 1.5 and 2 m to the row and positioned towards the trunk of a single plant. At this distance, the horizontal FoV varies between 2.3 and 3.1 m; thus, every image frames multiple plants of the row. Between the acquired images, 18 were captured from the east row and the other 18 from the west row in a time range between 8:29 and 8:43 in the morning (local time). These two parts of the dataset were equally divided into 9 images referring to plants before trimming and 9 to plants after trimming. As shown in Fig. 3(a), the acquisitions of the east row were backlit and less precise than those obtained from the west.

Specifically, the depth maps had smaller resolutions and narrower FoV with respect to the color images, as shown in Table 1. However, the built-in methods of the Azure Kinect Sensor SDK (v1.4.1) (Microsoft, 2022) allowed image transformation between color and depth geometry to map all pixels in the color reference to pixels of the depth map, or, equivalently, to the vertex of the corresponding point cloud (see Fig. 5).

The dataset includes the ground truth masks of the images needed for the evaluation of the segmentation quality. These masks were obtained by manual labeling considering three classes: leaves, bunches and background. As shown in Fig. 6, each pixel belonging to leaves, bunches and background was colored in green, blue and black, respectively. Considering the approach outlined in Casado-García et al. (2023), it was decided to incorporate also the grape class despite the primary goal of the work being the semantic segmentation of biomass volume rather than individual plant component segmentation. This decision was derived from experimental observations, which indicate that the three classes did not significantly impact the results. In addition, it allowed us to explore whether our proposed approach could effectively generalize the semantic classification when considering a different dataset acquired with a different sensor on a different cultivar.

[1] Open dataset available in Casado-García et al. (2023).
[2] Open dataset available in ISP STIIMA (2023).

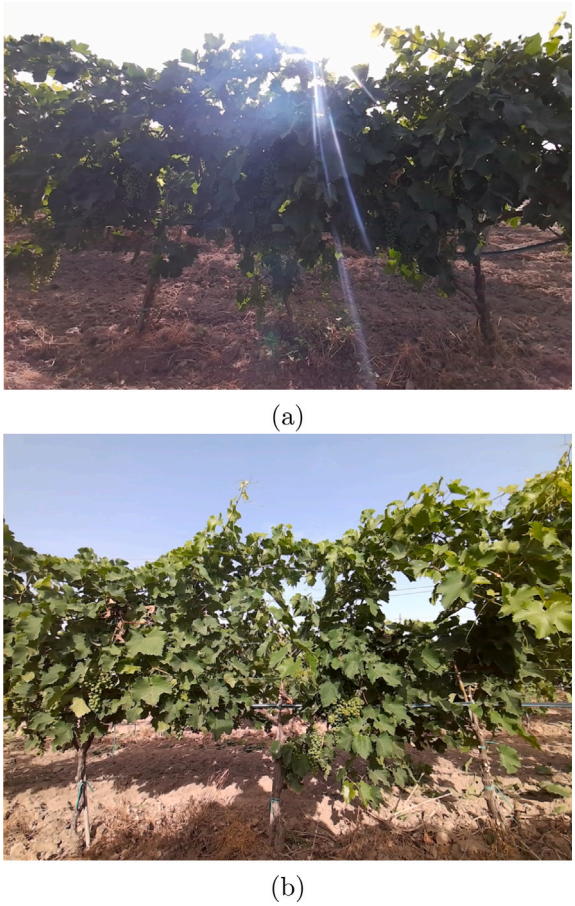

Fig. 3. Color images acquired with Microsoft Azure Kinect RGB-D camera of (a) the east and (b) west rows.

Table 1
Main parameters of the color and depth sensor of the Microsoft Kinect Azure camera.

	Depth sensor	Color sensor
Resolution (pixels)	640 × 576	2048 × 1536
FoV (H × V)	75° × 65°	90° × 74.3°

At the end of the trimming process, the removed biomass was collected and weighted after about 5 h (temperature of 29 °C - 31 °C, humidity of 45.4%–51.5% and atmospheric pressure of 1012 hPa–1014 hPa). Specifically, the measured mass of the removed leaves from the east and west rows was equal to 283 g and 244 g, respectively. This information will be the ground truth for the next point cloud analyses.

2.2. Deep neural networks

Fig. 7 shows the block diagram containing the steps for processing multi-modal data for semantic segmentation steps. In detail, the color images were first processed to recognize the biomass using deep learning segmentation architectures and mixed with the corresponding depth maps to produce the biomass point clouds.

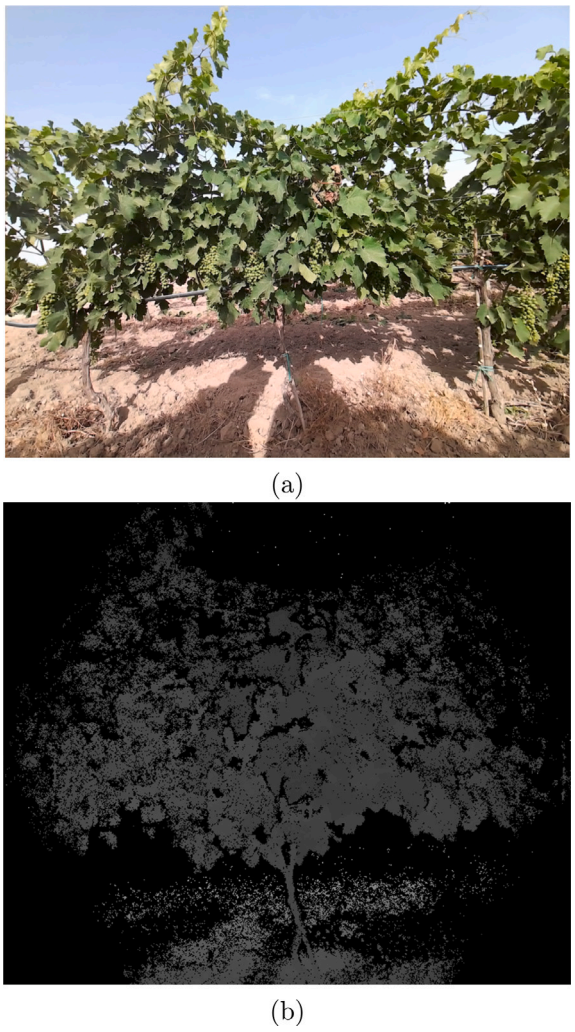

Fig. 4. Color (a) and depth images (b) produced by the Microsoft Azure Kinect camera.

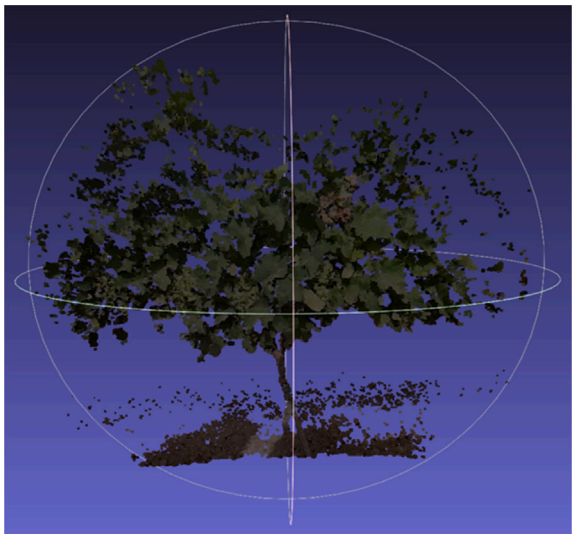

Fig. 5. Point cloud of the whole plant produced by the Microsoft Azure Kinect camera from the data in Fig. 4.

Table 2
Segmentation architectures and backbones presented in this work.

Segmentation architecture	Backbone
U-Net	ResNet50, ResNeXt50
DeepLabV3+	ResNet50, ResNeXt50
MANet	ResNet50, ResNeXt50

Several encoder–decoder neural networks for color image segmentation were compared for biomass analysis. Convolutional Neural Networks (CNNs) have become popular for semantic segmentation tasks. In these networks, the encoder, also known as the backbone, receives the input image to generate down-sampled feature maps and is composed of convolutional and max-pooling layers that gradually reduce the spatial resolution of the feature maps while increasing their depth; the decoder section generates a segmentation map by using upsampling blocks and convolutional layers that gradually increase the spatial resolution of the feature maps, restoring the size of the original image. The used architectures, shown in Table 2, were selected to achieve the best results in leaf segmentation, obtaining high accuracy in the shortest time, as demonstrated in Casado-Garcıa et al. (2022). The U-Net was the first network to propose an encoder–decoder architecture to perform semantic segmentation in medical contexts (Ronneberger et al., 2015). Its name derives from the U-shape assumed by the network itself; the encoder uses a typical CNN architecture and the decoder consists of up-convolutions and concatenations with features from the encoder path. The DeepLabV3+ is a network built upon the previous versions of the DeepLab architecture and incorporates atrous convolutions (Chen et al., 2018), to improve its accuracy and efficiency. Finally, the MANet (Multi-scale Attention Net) uses multiple attention mechanisms that help the network to focus on important regions of the input image and ignore irrelevant information (Li et al., 2021). The backbone can be of different types; in this work, ResNet50 and ResNeXt50 were considered. The first is a CNN where the "Res" in ResNet stands for "residual", which refers to the use of residual blocks in the architecture, and the 50 indicates the number of layers (He et al., 2016). Residual blocks can overcome the degradation problem due to vanishing gradients of the backpropagating loss function, which does not allow the network to learn. Residual blocks let the loss gradient flow directly towards the input, thus improving the learning of the model. The ResNeXt50 is a variant of the ResNet where the "X" stands for "Next dimension" which indicates an additional dimension, called cardinality, that refers to the size of the set of transformations, i.e. to the number of independent paths in a residual block (Xie et al., 2017).

Any implementation of the proposed networks followed the concept of transfer learning (Weiss et al., 2016), which involves using pre-trained models that have already learned useful functions from a large dataset. In the proposed approach, all the backbones were pre-trained by the ImageNet dataset (Deng et al., 2009). In this way, these pre-trained models are able to learn general image features that can be used in various computer vision tasks, including plant part recognition. The pre-trained networks were considered as a starting point to train the network on the new dataset, i.e. the Training Dataset provided by the RealSense D435, by making changes to the network weights. This process of fine tuning helps specialize the neural network for the new task, in this case the grapevine biomass recognition task, without completely losing the knowledge learned from the pre-trained models. The concept of transfer learning was also applied during the testing phase as a different dataset was used to evaluate the performance of the network. The use of a test dataset different from the training one was chosen to evaluate the ability of the neural network to generalize and see if it is able to recognize similar patterns even in data coming from a different sensor.

The network training requires some operations on the dataset and the selection of some parameters. First, data Augmentation (Negassi et al., 2022) involved adding more data to the existing dataset by

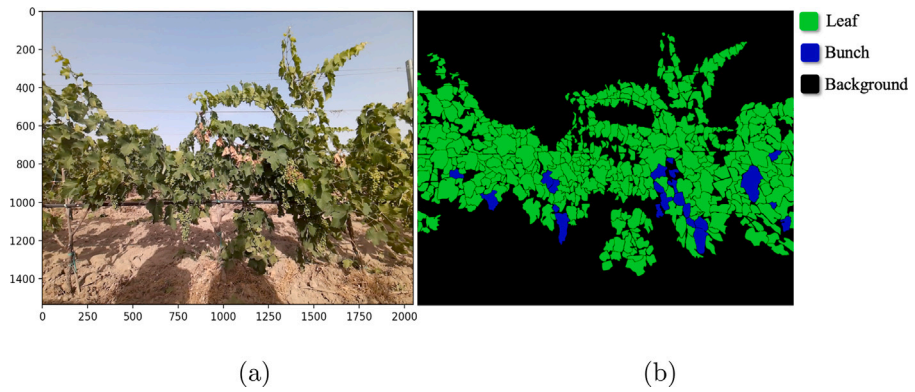

Fig. 6. (a) Input image acquired by the Azure Kinect camera and (b) corresponding image annotation with the three classes of interest. Color correspondence is shown in the legend.

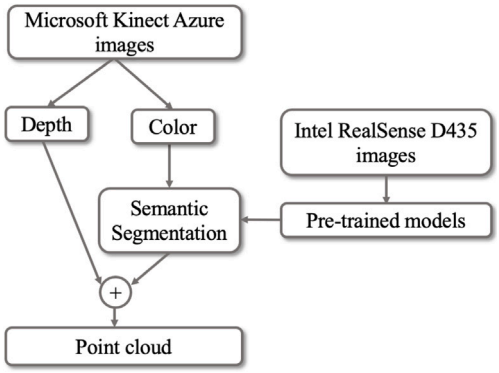

Fig. 7. Block diagram of the data processing for biomass segmentation.

applying different transformations to the images. The augmentation was implemented using the Albumentations library (Buslaev et al., 2020) in Python which allowed applying the flipping and transposition operations with a probability of 50%, the rotation operation with a probability of 40% and a rotation limit of 10 degrees, and the Grid Distortion operation. This approach increased the actual cardinality of the training dataset improving its variability for better generalization capabilities.

Subsequently, a resize operation on the input images was performed to change the original aspect ratio of the training images (9:16) and maintain the correspondence with images produced by the Azure Kinect (4:3).

Then, to avoid overfitting, a patience parameter was set; this parameter is related to the concept of Early stopping that allows suspending training when the performance on the validation data set starts to deteriorate (Prechelt, 2002). In this work, patience was set to 3, which means that the training stops if the validation loss does not improve for three consecutive epochs.

Further hyperparameters, such as the number of epochs, the batch size and the learning rate, were also set (Bengio, 2012). One of the most important is the learning rate as it decides how fast the network weights are tuned to find the loss minimum. A small learning rate makes the model converge slowly, while a large learning rate makes the model diverge. The solution is to find the optimal compromise for the learning rate related to the considered data. Using the Fastai library, it is possible to perform a preliminary experiment to find a reasonable learning rate for training the model (Gugger, 2018b). In this work, different learning rate values were chosen using the *1-cycle learning rate policy*, where the learning rate was gradually increased to a maximum and then reduced during training to improve both the speed and accuracy of the training (Gugger, 2018a). Moreover, according to the transfer learning approach, two training phases were implemented: in the first, only the last layers were trained starting with a high learning rate (before unfreeze); in the second, the whole network fine-tuned its weights with a lower learning rate (after unfreeze). All the hyperparameters are reported in Table 3

Table 3
Training hyperparameters for all the segmentation networks.

Parameters	Values
Patience	3
Epoch	20
Batch size	4
Learning rate (before unfreeze)	$10^{-4}, 10^{-3}$
Learning rate (after unfreeze)	10^{-7}

The mean segmentation accuracy (MSA) was considered to evaluate the training results. The MSA is computed as the average value of the accuracy of each class.

Subsequently, after the training phase, at the end of the testing phase the output segmentation masks were compared with ground truth images using suitable evaluation metrics. In a *One-vs-All* approach, given a class, T_P, T_N, F_P, F_N are the number of True Positive, True Negative, False Positive and False Negative predictions of the class against all the others, respectively. In this work, the following metrics were considered (Grandini et al., 2020; Fränti and Mariescu-Istodor, 2023; Roy and Ameer, 2021; Csurka et al., 2013):

1. *Accuracy*: it measures the percentage of correctly classified pixels in the entire image, regardless of class:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{1}$$

2. *Mean Segmentation Recall (MSR)*: it is the average of the recall values of each class, i.e. sensitivity in labeling pixels, computed as:

$$Recall = \frac{T_P}{T_P + F_N} \tag{2}$$

3. *Intersection over Union score (IoU score)*: it follows the definition of the Jaccard index, which calculates the ratio of the intersection between the predicted and ground truth images to their union. Its formulation is given by:

$$IoU = \frac{T_P}{T_P + F_N + F_P} \tag{3}$$

In the proposed experiments, two IoU scores were used:

- *Mean IoU score*: it measures the average percentage of correctly classified pixels across all classes in the image;

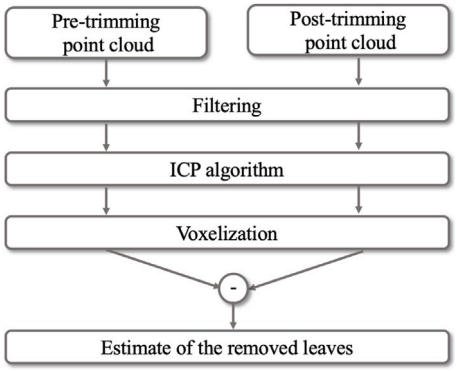

Fig. 8. Block diagram of the point cloud comparison for estimating the removed biomass after trimming.

- *Weighted IoU score*: it is the average IoU of all classes, weighted by the number of pixels of the classes.

4. *BF Score (Boundary F1 Score)*: it measures the F1 score of the boundary pixels between the predicted and ground truth images. For each image, the Mean BF Score is the average BF score of all classes in that image.

2.3. Point cloud processing

Fig. 8 describes the point cloud processing steps carried out to estimate the removed biomass, starting from the data acquired before and after the trimming phases. A point cloud is a collection of data points mapped in three dimensions (Bello et al., 2020) that provides a detailed, three-dimensional representation of the crops. The point cloud processing in Fig. 8 aims to find a match between the estimated volume of the removed leaves and their actual mass (ground truth).

As a first step, the couples of point clouds of corresponding plants before and after trimming were registered. This was an important step to estimate the optimal rigid transformation between two point clouds since each acquisition had its own pose. The ICP (Iterative Closest Point) method proposed by Besl and Mckay in 1992 (Besl and McKay, 1992) is the most widely used algorithm for point cloud registration. Starting from two point clouds (a template and a reference), pairs of corresponding points are iteratively extracted. A rigid transformation is applied to the template point cloud to reduce the distance between the couples of points. The algorithm ends with an optimal transformation that minimizes the distance measurement between the two point clouds.

The filtering operation was also crucial because point clouds produced by low-cost sensors suffer from noise sources and contain outliers due to sensor limitations, ambient lighting, and target reflectance, which produce artifacts in the scene. Outliers were removed from the point cloud by clustering vertices with a k-nearest neighbor algorithm. The average distance to the neighbors of all vertices was thus computed and compared to an outlier threshold value. Vertices having average distances above the specified threshold were removed from the point cloud.

The final step in point cloud processing was voxelization. The 3D space was converted into a voxel representation through a resampling process, equivalent to the pixel discretization in 2D images (Xu et al., 2021). Each voxel was set to an occupancy state if it enclosed a vertex of the point cloud. In this way, the resulting space became ordered and allowed the final comparison of the spatial distribution of biomass volumes.

Table 4
MSA results and time required in training all the segmentation architectures. The best results are in bold.

Segmentation architecture		MSA (%)	Time (min)
U-Net	ResNet50	92.55	50.42
	ResNeXt50	92.03	40.68
DeepLabV3+	ResNet50	92.56	16.55
	ResNeXt50	94.00	**12.05**
MANet	ResNet50	91.02	59.50
	ResNeXt50	**94.40**	113.13

3. Results

In the following, the results of image segmentation, both during the training and test phases, and the point cloud processing are reported. It is worth noticing that the training and testing phases of the proposed neural networks were performed on different datasets obtained with different sensors. Therefore the data processing procedures and the parameter settings are described in detail.

In the next experiments, all the networks are trained using the PyTorch (Paszke et al., 2019) and FastAI (Howard and Gugger, 2020) libraries on a consumer workstation (Intel NUC 9 Pro Kit), equipped with an Nvidia GeForce GTX 1660 SUPER GPU.

3.1. Training results

As stated in Section 2.1.1, the dataset of 315 images acquired by the Intel RealSense D435 sensor was used for training.

The values in Table 4 were achieved using the Shuffle function in order to mix the dataset at each epoch and prevent the model from learning the order of the data. The MANet with ResNeXt50 backbone performed the best training as it reached the highest value of MSA. Table 4 also shows the time required for training each network. In this case, the DeepLabV3+ with ResNeXt50 backbone achieved the shortest time, with comparable MSA (94.00%) as the MANet with ResNeXt50 backbone.

The images in Fig. 9 compare the ground truth labels with the predicted masks. The MANet network with ResNeXt50 backbone produced good results since it was able to well-segment both the plant canopy and the grape bunches. On the contrary, the MANet network with the ResNet50 backbone was the worst in bunch segmentation, thus returning the lowest MSA value. However, all the networks produced robust results in segmenting the canopy well, while having problems identifying the bunches, which are often included in the leaf class.

3.2. Testing results

After the training phase, the networks were tested on the actual dataset made of the Azure Kinect images to produce predictions.

For an initial evaluation of the results, Fig. 10 shows two sets of segmentation results obtained by processing images of the west and east vineyard rows. The masks predicted from the west row were more precise than those from the east row. Specifically, in many images from the east row, the networks were not able to identify the central part of the plants due to the presence of sun rays. From these images, only the DeepLabV3+ network with ResNet50 backbone made a discrete prediction. On the contrary, all the networks provided good results in segmenting the images of the west row. In particular, the best segmentation was obtained by the U-Net with ResNet50 backbone, while the predictions produced by MANet were the worst regardless of the backbone. Furthermore, the networks were not able to correctly segment the grape bunches in both west and east images. This aspect was due to the difference in the cultivars of the training and test datasets. Although the plant canopies were comparable, the appearance of the grape bunches had significant differences that led to misdetections during testing.

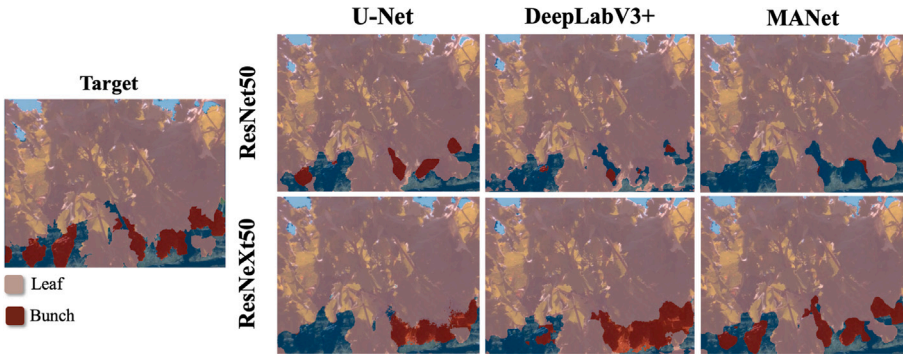

Fig. 9. Predictions on a sample image of the training dataset. The target image is the reference ground truth.

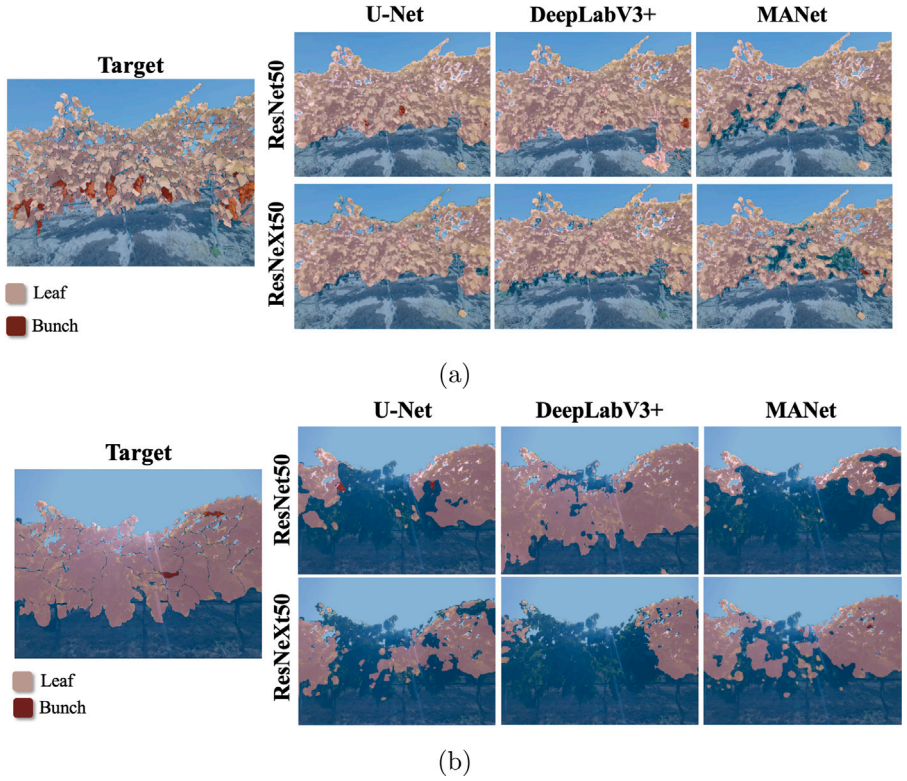

Fig. 10. Predicted segmentation masks for test images of (a) the west and (b) the east rows, computed by the proposed segmentation architectures.

These qualitative results can be verified by quantitative metrics. Table 5 and Table 6 report all the outcomes of the proposed networks for the west and east rows, respectively. Considering the metrics from the west row in Table 5, the results obtained with all the networks are satisfactory since the segmentation masks are generated by training and testing the networks on images produced by two different sensors, which introduced variability and made the segmentation task more challenging. Specifically, the best network is the U-Net with the ResNet50 backbone which reached an accuracy of 92.10%, an MSR of 92.21% and a mean IoU score of 85.09%. In all cases, the BF score was the metric with the lowest values but, on the other hand, the IoU presented good results. It suggests that the model was good at accurately predicting the overall shape of the segmented regions, as reflected by the high IoU, but it could not accurately delineate the boundaries of those regions, as reflected by the low BF score. However, a low Mean BF Score was not a significant problem since retrieving the leaf boundaries is not crucial to this work, which aimed at segmenting the leaf mass as a whole. In this regards, the high IoU score indicates that the algorithm was performing well. With reference to the east row, the network that made the best prediction was the DeepLabV3+ with the ResNet50 backbone (Accuracy, MSR and IoU of 79.45%, 78.68% and 65.00%, respectively) accordingly with the results in Fig. 10. In this case, as expected, even the best accuracy was significantly lower than that from the west row, because of the poor quality of the input image due to direct sunlight.

3.3. Point clouds results

The point cloud analysis gives information about the biomass removed during the trimming process. In this context, only the plant leaves structure, extracted from both ground truth and segmentation masks network-predicted, was considered (see Fig. 11). Both masks were used to check if predictions were accurate enough to compute the removed leaf volume, compared to the ground truth.

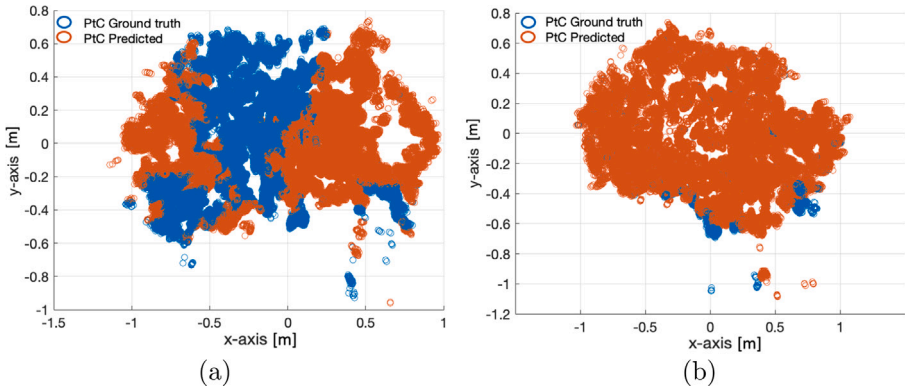

Fig. 11. Leaf mass point clouds for (a) the east and (b) the west rows. Blue circles refer to the point clouds extracted from the ground truth masks, while orange circles refer to the points computed from the predicted segmentation masks.

Table 5
Segmentation results for the images of the west row. The results in bold are the best.

Segmentation architecture		Accuracy (%)	MSR (%)	Mean IoU Score (%)	Weighted IoU Score (%)	Mean BF Score (%)
U-Net	ResNet50	**92.10**	**92.21**	**85.09**	**85.41**	**73.98**
	ResNeXt50	89.08	88.37	79.73	80.29	69.64
DeepLabV3+	ResNet50	87.14	85.35	76.05	76.92	68.50
	ResNeXt50	90.25	89.80	81.78	82.25	70.74
MANet	ResNet50	90.61	90.14	82.37	82.83	73.0
	ResNeXt50	88.90	87.75	79.26	79.91	73.35

Table 6
Segmentation results for the images of the east row. The results in bold are the best.

Segmentation architecture		Accuracy (%)	MSR (%)	Mean IoU Score (%)	Weighted IoU Score (%)	Mean BF Score(%)
U-Net	ResNet50	65.67	63.97	44.56	45.33	50.15
	ResNeXt50	58.24	56.10	34.10	35.16	33.99
DeepLabV3+	ResNet50	**79.45**	**78.68**	**65**	**65.28**	**57.30**
	ResNeXt50	58.51	56.38	34.46	35.51	40.34
MANet	ResNet50	60.37	58.39	37.30	38.25	51.79
	ResNeXt50	73.45	72.19	55.68	56.19	55.80

As a first step, it is necessary to compare the effects of the use of segmentation predictions on the point clouds of the biomass. Fig. 11 shows a superposition of the same point clouds obtained considering the ground truth masks (blue circles) and the corresponding predictions (orange circles). In this case, considering the west or the east rows produced different results. Specifically, although the point clouds of the west row in Fig. 11(b) were almost perfectly overlapped, the comparison of biomass point clouds from the east row in Fig. 11(a) revealed a large void in the point cloud obtained from segmentation predictions. This incorrect superposition is clearly due to the wrong segmentation of the biomass, especially in the center, because of the direct sunlight affecting the east row data.

After filtering the point clouds, the ICP algorithm was applied considering as 'template' and 'reference' the point clouds acquired before and after trimming, respectively. It is important to notice that the ICP algorithm was applied to the whole point clouds of the whole plant scene, as provided directly by the Azure Kinect camera, including also trunks, soil and artificial infrastructures, such as poles. In this way, the ICP registration was more robust as fixed objects were considered in contrast to the plant canopy which changed its shape because of the trimming procedure.

Observing Fig. 12, the trunks and the poles, if any, were aligned, but the leaf structure, especially at the edges, i.e. on the sides, was not perfectly registered. This depends on the different points of view of the camera across the acquisitions, that framed different areas of the same plants. In addition, uncontrollable atmospheric agents, such as the wind, caused intrinsic differences between the acquired images.

Table 7
Comparison between mass, measured on field, and computed volumes obtained using the ground truth segmentation masks for the east and west rows.

	East	West
Removed mass (from ground truth) [g]	283	244
Estimated volume changes [m^3]	0.457	0.409

On the other hand, the ICP algorithm performed well for both the west and east rows, since the direct light of the east rows did not affect the depth maps and, thus, the point clouds, unlike the color ones.

The transformation matrix from the ICP algorithm was used to register the point clouds. As shown in Fig. 13, referring to a couple of data from the west row, there was not a perfect overlap. This alteration included all the previously mentioned sources (noise, implicit alterations, leaves movements, etc.), but was mainly ascribable to the trimming process: the removed leaves determined a difference between the two point clouds. This was even more evident by comparing the color images before (Fig. 14(a)) and after trimming (Fig. 14(b)), where the reduction in the leaf mass in the red circles is also present in the volume highlighted in Fig. 14.

After the transformation, the voxelization sampled the 3D space into a grid cell with side lengths of 0.1 m. Fig. 15 shows an overlap of the voxels enclosing vertices of the pre- and post-trimming point clouds, except in those areas (in the red circle) where the leaves were trimmed.

Finally, a differential operation was carried out to highlight the voxels of the pre-trimming point clouds that were not present in the post-trimming ones, assuming that these elements correspond to the removed leaves. At the step, further filtering was applied to eliminate those resulting voxels at the highest depth, i.e. with high z coordinates. These detected differences should not be related to the trimming procedure since it can remove only the most superficial leaves, at lower z coordinates. The result of this operation is shown in Fig. 16, where the points marked in red correspond to the removed volumes, as seen in Fig. 13.

The described processing was performed on all the point clouds in the east and west rows. To prove the validity of the pipeline itself, the first experiments were run considering the ground truth segmentation maps, instead of the actual predictions.

Each detected voxel of alteration was associated with a volume of 10^{-3} m^3. Since the measured mass of the removed leaves of the west row was lower, the volume should also be smaller than that of the east row. The hypothesis was confirmed by the results reported in Table 7.

To evaluate the consistency of the results, the density of the detected removed biomass was computed from both the east and west rows as the ratio between the on-field measured mass and the estimated

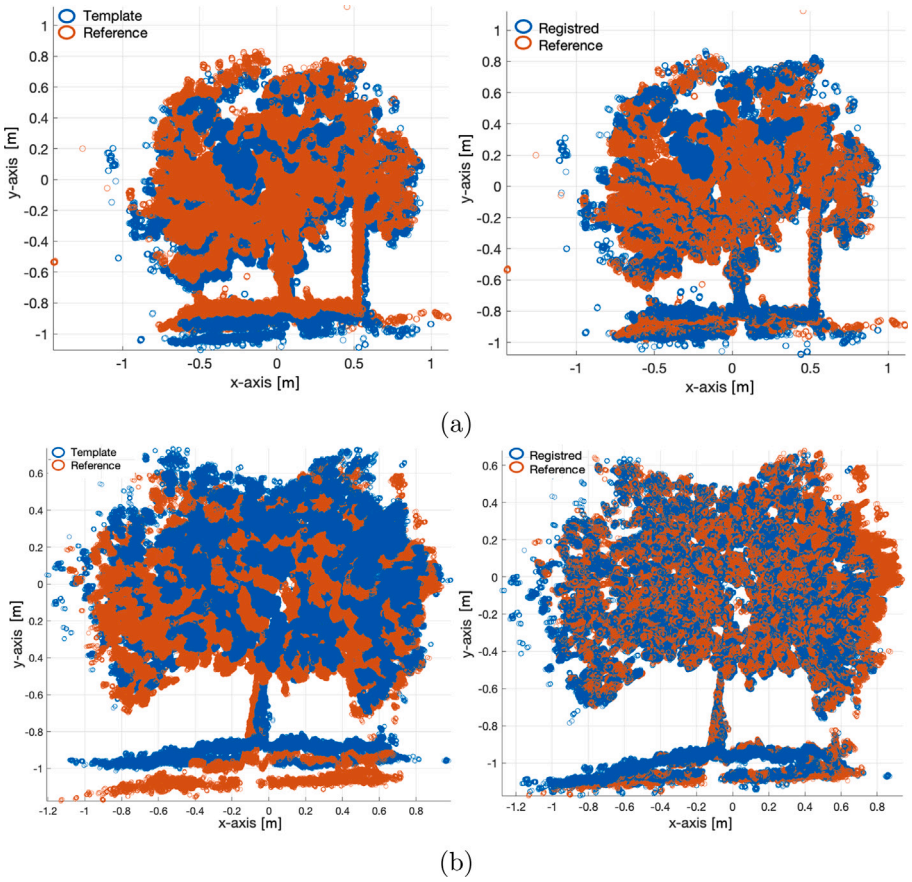

Fig. 12. Comparison between the pre- and post-trimming point clouds before and after the application of the ICP algorithm, for (a) the west and (b) east row. The pre-trimming point clouds are the 'template' and the post-trimming ones are the 'reference'.

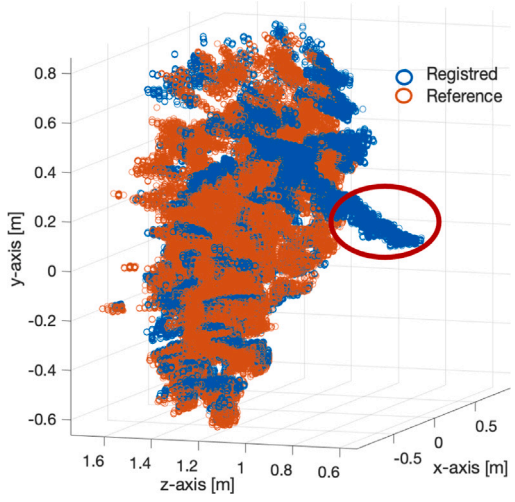

Fig. 13. Overlay of the pre- and post-trimming point clouds from the images in Fig. 14. The red circle encloses the removed leaves.

volume. These two values should be ideally equal if the volumes are perfectly estimated. The density values were equal to 619.3 g/m^3 and 596.6 g/m^3 for the east and west rows, respectively. The low difference of 22.7 g/m^3 in the estimated density values was mainly due to residual noise that may still be present despite the processing, as well as misalignment between the point clouds captured before and after trimming.

Table 8
Comparison between mass, measured on field, and computed volumes obtained using the predictions from the DeeplabV3+ with ResNet50 backbone for the east and west rows.

	East	West
Removed mass (from ground truth) [g]	283	244
Estimated volume changes [m^3]	0.388	0.431

Once the point cloud processing pipeline is verified, it is possible to merge the results of the segmentation predictions made by the proposed neural networks. The first evaluation was made considering the DeeplabV3+ with ResNet50 backbone, which produced acceptable results even for the east row. However, from the results in Table 8, the estimated volume of the removed biomass of the west row was higher than that of the east row, contrary to expectations. This result was clearly biased by the wrong segmentation of the images of the east rows, whose biomass was underestimated. On the contrary, the volume obtained for the west row was comparable with that obtained by using the ground truth masks, thus highlighting the quality of the segmentation, even if the training was run on a dataset of different specifications.

The results of the application of the whole pipeline on the images of the west row are in Table 9. Considering the accuracy results, the U-Net with ResNet50 backbone, which showed the best segmentation results, was expected to estimate the closest volume changes in comparison to that in Table 7, obtained from the ground truth masks and equal to 0.409 m^3. Similarly, the DeepLabV3+ with ResNet50 backbone was expected to produce the least accurate results. On the contrary, Table 9

(a) (b)

Fig. 14. Comparison between color images (a) before and (b) after trimming. The red circle encloses the removed leaves.

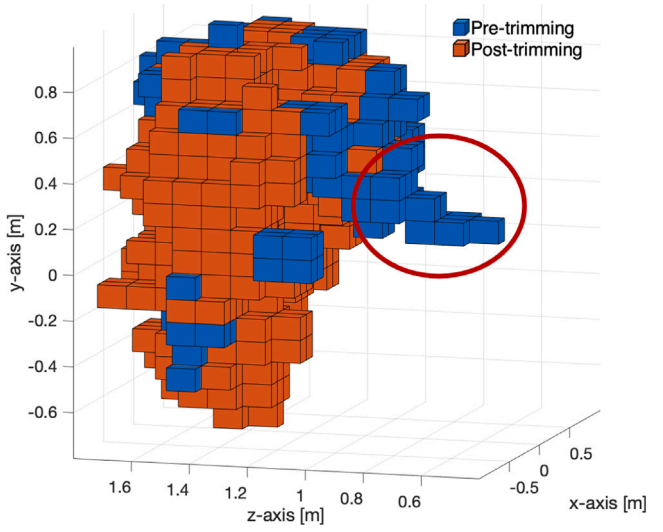

Fig. 15. Voxelization of the pre- and post-trimming point clouds (blue and orange cubes, respectively). The red circle encloses the area where the leaves were removed.

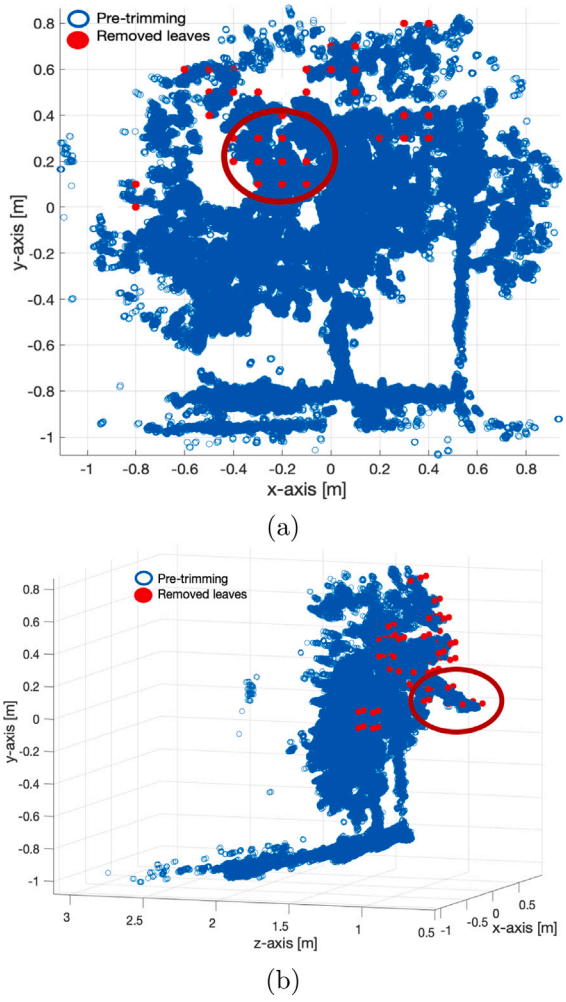

Fig. 16. Representation of the removed leaves (red marks) on the pre-trimmed point cloud from two points of view; the red circle encloses the area where leaves were removed.

shows that the network that produced the closest result to 0.409 m^3 was the DeepLabV3+ with ResNext50 backbone, while the worst was the MANet with ResNeXt50 backbone.

Keeping the ResNeXt50 backbone, the same trend as the segmentation metrics was observed: the best architecture was the DeepLabV3+, followed by U-Net and MANet. The results followed a different trend when ResNet50 was used as the backbone. In this case, the best performing network was the DeepLabV3+, which however scored the worst accuracy value in Table 5.

A further comparison could be made by keeping the segmentation architectures and varying the backbone. Observing the accuracy values in Table 5, the best backbone for the U-Net and the MANet was the ResNet50, while the DeepLabV3+ performed better with the ResNeXt50. The same trend can be observed considering the results obtained for estimating the removed biomass volume.

4. Discussions

The primary objective of the current work is to monitor the trimming process to analyze the canopy volume of vine plants. To our knowledge, there is a lack of studies in this field. There are some works in which the canopy of other plants has been analyzed, but the volume estimation in a vineyard was not considered. The methodologies proposed in these studies do not combine RGB or three-dimensional image analysis and often rely on data acquired by unmanned aerial vehicles (UAVs). In contrast, our study adopts a ground-based platform, which allows the capture of more representative images, showing the complex leaf structure of grapevine plants in more detail. In Section 3.1, it was found that all the proposed deep neural networks demonstrated a remarkable ability to segment the canopy accurately. During the training phase, the outcomes obtained from these networks demonstrated comparability in terms of Mean Segmentation Accuracy (MSA),

Table 9
Volume and accuracy values obtained using the predicted segmentation masks produced by the proposed networks for the west side. Values close to 0.409 m^3 are the best.

Segmentation architecture		Estimated volume [m^3]	Accuracy(%)
U-Net	ResNet50	0.472	**92.10**
	ResNeXt50	0.500	89.08
DeepLabV3+	ResNet50	0.431	87.14
	ResNeXt50	**0.421**	90.25
MANet	ResNet50	0.477	90.61
	ResNeXt50	0.519	88.90

achieving values higher than 90%. The time required for each network to converge and achieve the optimal segmentation result varied significantly among the different networks. However, time differences could not be considered particularly significant; since the training is performed only once, investing a long time in this part of the work does not pose any significant issues. Another aspect that emerged during the training phase was the problem of accurately identifying the grape bunches. Nevertheless, this was not crucial since the goal was to segment the biomass volume semantically rather than performing an instance segmentation of the individual components of the plants.

Also considering the testing results (Section 3.2), the networks could not detect the grapes; as previously stated, this was irrelevant since bunch pixels are labeled as leaves not altering the estimation of the whole biomass volume. The canopy segmentation achieved by all the networks was particularly satisfactory considering that the training and testing phases involved images captured by two different cameras. This introduction of variability between the camera sources increased the complexity of the segmentation task. Despite this challenge, the results obtained in both phases demonstrated agreement, emphasizing the efficacy of the transfer learning approach. This approach, initially inspired by the work of Casado-García et al. (2023), has further emphasized the importance of the transfer learning procedure as it proved advantages in overcoming the limited availability of data and reducing the labor-intensive process of manually labeling many images. It makes the use of deep learning methods less time- and effort-consuming.

In the last part of the work, the point clouds of the plant canopy were analyzed to estimate potential volume changes resulting from the trimming process. First of all the point clouds were not influenced by the sunrays, as computed at short distances, following the principle of ToF in the infrared spectrum. In this case, there may be other problems, such as atmospheric agents (i.e., wind), that could compromise the application of the ICP algorithm. In fact, further works will be aimed at improving the registration of the point clouds by giving more emphasis to the fixed structures of the scene, such as trunks, poles and soil, even segmented by preliminary deep neural networks. This will also enable the effective comparison of point clouds acquired at different phenological phases.

Furthermore, the analysis of the results proved that the network that achieved the best segmentation results did not necessarily provide the most accurate estimate of the removed leaves. The DeepLabV3+ with both the ResNet50 and ResNeXt50 backbones provided the best estimation of removed biomass volume, since these values are the closest to the reference value of 0.409 m^3, achieved with the ground truth masks. This was in contrast with the accuracy of the segmentation process, where the U-Net with ResNet50 showed the best results. Despite U-Net's superior overall segmentation capability, it struggled to accurately identify areas of the canopy that were most affected by the trimming operation, such as extended branches. Conversely, the DeepLabV3+ network with ResNeXt50 improved performance in evaluating the removed volume, indicating a greater ability to segment the specific regions affected by the trimming process. These volumes were much more significant in the proposed work, which aimed at estimating the removed biomass instead of straight canopy segmentation. Probably, this is due to the U-Net architecture that consider the canopy as a whole, where every part had the same semantic weight, i.e. significance. On the other hand, the atrous convolutions, used by the DeepLabV3+, allows expanding the network's field of view without increasing the number of parameters, enabling the extraction of information from a broader area of the input without enlarging the size of the filters. Having a broader receptive field can better model extended canopy parts, such as long branches, which are typically subjected to the trimming process.

This discrepancy in the results emphasizes the significance of considering not only the overall quality of the segmentation but also the network's ability to identify the most relevant areas of interest for the specific analysis. In this case, only after the accurate evaluation of all the processing steps carried out on color images and point clouds is it possible to select the best-performing network.

5. Conclusions

This paper presents a framework for evaluating the canopy volume changes of vine plants after a trimming process, analyzing images obtained during on-field experiments with a low-cost RGB-D camera, namely the Microsoft Azure Kinect. The main findings of the proposed paper were: to demonstrate that the transfer learning approach allows the generalization of semantic segmentation on datasets different from those used in the learning phase; to verify that the use of an initial semantic segmentation of the canopy guarantees an effective estimation of the plant volumes; to confirm that the point cloud registration allows volume change evaluation after the trimming process and the portability of the method on any mobile platform.

Several pre-trained deep neural networks were compared, trained on 315 images of a publicly available dataset, and tested on 36 natural images. The latter were acquired from two grapevine rows (from both west and east sides) of a cultivar different from the training one and under different sunlight conditions. In particular, the segmentation networks (U-Net, DeepLabV3+ and MANet architectures, with ResNet or ResNeXt backbone), pre-trained by the ImageNet dataset, were tuned using images of Intel RealSense D435 dataset, to take advantage of existing manual annotation. Segmentation metrics on the test set proved the network's ability in model generalization, especially on the images of the west row. In this case, the U-Net with ResNet backbone scored 92.10% and 85.09% in accuracy and IoU. Conversely, the presence of direct sunlight on the images of the east row downed the quality of the results.

The second part of the study analyzed the point clouds of the plant canopy to estimate possible volume changes due to an actual trimming process. A specific framework, including point cloud filtering, registration and voxelization, was presented to compare, i.e. differentiate, 3D models and estimate the removed biomass. The first analyses used the ground truth segmentation masks as input to the point cloud processing, demonstrating its effectiveness regardless of the quality of the preliminary segmentation. The results showed a correct correlation with the weights of the cut leaves for both rows, east and west. On the other hand, the outcomes of the segmentation networks produced three results: (i) the best agreement of volume estimation (0.421 m^3) with the expected ground truth (0.409 m^3) was achieved by the DeepLabV3+ with a ResNext50 backbone, which was not the best network in segmentation metrics. (ii) the DeepLabV3+ with a ResNext50 backbone was the most efficient network in training, with the shortest training time (about 12 min); (iii) the misclassification of the images of the east row, due to direct sunlight, produced an underestimation of the volume of the removed biomass. Future work will focus on further experiments by acquiring larger datasets in challenging lighting conditions to avoid the misclassifications that could have affected the actual performances.

In conclusion, this work demonstrated that multimodal RGB-D cameras allow the analysis of the plant canopy. By mounting a low-cost sensor on the equipment for routine vineyard maintenance, such as tractors, farmers can have a tool for monitoring the growth of the canopy over time, maintaining the crop yield and minimizing potential damages that may affect the harvest.

CRediT authorship contribution statement

A. Bono: Writing – original draft, Conceptualization, Methodology. **R. Marani:** Data curation, Methodology, Writing – review & editing, Conceptualization, Software. **C. Guaragnella:** Supervision. **T. D'Orazio:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data have been uploaded online for public sharing with the scientific community.

Acknowledgments

This paper was funded by the "TEBAKA - TErritorial BAsic Knowledge Acquisition" research project (MUR PON Agrifood Program, Grant No. ARS01_00815). The authors would like to thank the "Conte Spagnoletti Zeuli" farm in Andria (Italy) for hosting the experiments and Mr. Michele Attolico for his crucial technical support.

References

Andújar, D., Moreno, H., Bengochea-Guevara, J.M., de Castro, A., Ribeiro, A., 2019. Aerial imagery or on-ground detection? An economic analysis for vineyard crops. Comput. Electron. Agric. 157, 351–358. http://dx.doi.org/10.1016/j.compag.2019.01.007.

Apostolidis, K.D., Kalampokas, T., Pachidis, T.P., Kaburlasos, V.G., 2022. Grapevine plant image dataset for pruning. Data 7 (8), http://dx.doi.org/10.3390/data7080110.

Arnó Satorra, J., Martínez-Casasnovas, J.A., Ribes-Dasil, M., Rosell, J.R., 2009. Precision viticulture. Research topics, challenges and opportunities in site-specific vineyard management. Span. J. Agric. Res. 7 (4), 779–790. http://dx.doi.org/10.5424/sjar/2009074-1092.

Bello, S.A., Yu, S., Wang, C., Adam, J.M., Li, J., 2020. Deep learning on 3D point clouds. Remote Sens. 12 (11), 1729. http://dx.doi.org/10.3390/rs12111729.

Bengio, Y., 2012. Practical recommendations for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade, second Ed. Springer, pp. 437–478. http://dx.doi.org/10.48550/arXiv.1206.5533.

Besl, P.J., McKay, N.D., 1992. Method for registration of 3-D shapes. Sensor fusion IV: Control paradigms and data structures. Int. Soc. Opt. Photonics 1611, 586–606. http://dx.doi.org/10.1117/12.57955.

Blekos, A., Chatzis, K., Kotaidou, M., Chatzis, T., Solachidis, V., Konstantinidis, D., Dimitropoulos, K., 2023. A grape dataset for instance segmentation and maturity estimation. Agronomy 13 (8), 1995. http://dx.doi.org/10.3390/agronomy13081995.

Borgogno-Mondino, E., Lessio, A., Tarricone, L., Novello, V., de Palma, L., 2018. A comparison between multispectral aerial and satellite imagery in precision viticulture. Precis. Agric. 19, 195–217. http://dx.doi.org/10.1007/s11119-017-9510-0.

Botta, A., Cavallone, P., Baglieri, L., Colucci, G., Tagliavini, L., Quaglia, G., 2022. A review of robots, perception, and tasks in precision agriculture. Appl. Mech. 3 (3), 830–854. http://dx.doi.org/10.3390/applmech3030049.

Bramley, R.G.V., Hamilton, R.P., 2004. Understanding variability in winegrape production systems: 1. Within vineyard variation in yield over several vintages. Aust. J. Grape Wine Res. 10 (1), 32–45. http://dx.doi.org/10.1111/j.1755-0238.2004.tb00006.x.

Buslaev, A., Iglovikov, V., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A., 2020. Albumentations: Fast and flexible image augmentations. Information 11 (2), 125. http://dx.doi.org/10.3390/info11020125.

Casado-Garcıa, A., Heras, J., Milella, A., Marani, R., 2022. Semi-supervised deep learning and low-cost cameras for the semantic segmentation of natural images in viticulture. 23, pp. 1–24. http://dx.doi.org/10.1007/s11119-022-09929-9,

Casado-García, A., Heras, J., Milella, A., Marani, R., 2023. Generalization of deep learning models applied to semantic segmentation of in-field natural images in vineyards. http://dx.doi.org/10.3920/978-90-8686-947-3, URL https://github.com/ispstiima/ECSDVineyardDataset.git.

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818. http://dx.doi.org/10.48550/arXiv.1802.02611.

Cisternas, I., Velásquez, I., Caro, A., Rodríguez, A., 2020. Systematic literature review of implementations of precision agriculture. Comput. Electron. Agric. 176, 105626. http://dx.doi.org/10.1016/j.compag.2020.105626.

Comba, L., Biglia, A., Aimonino, D.R., Gay, P., 2018. Unsupervised detection of vineyards by 3D point-cloud UAV photogrammetry for precision agriculture. Comput. Electron. Agric. 155, 84–95. http://dx.doi.org/10.1016/j.compag.2018.10.005.

Csurka, G., Larlus, D., Perronnin, F., Meylan, F., 2013. What is a good evaluation measure for semantic segmentation? In: Bmvc, Vol. 27. Bristol, pp. 10–5244. http://dx.doi.org/10.5244/C.27.32.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Di Gennaro, S.F., Matese, A., 2020. Evaluation of novel precision viticulture tool for canopy biomass estimation and missing plant detection based on 2.5 D and 3D approaches using RGB images acquired by UAV platform. Comput. Electron. Agric. 16, 1–12. http://dx.doi.org/10.1186/s13007-020-00632-2.

Fränti, P., Mariescu-Istodor, R., 2023. Soft precision and recall. Pattern Recognit. Lett. 167, 115–121. http://dx.doi.org/10.1016/j.patrec.2023.02.005.

Grandini, M., Bagli, E., Visani, G., 2020. Metrics for multi-class classification: an overview. http://dx.doi.org/10.48550/arXiv.2008.05756, arXiv preprint arXiv:2008.05756.

Gugger, S., 2018a. The 1cycle policy. URL https://sgugger.github.io/the-1cycle-policy.html?ref=derekchia.com.

Gugger, S., 2018b. How do you find a good learning rate. URL https://sgugger.github.io/how-do-you-find-a-good-learning-rate.html.

Hall, A., Lamb, D.W., Holzapfel, B., Louis, J., 2002. Optical remote sensing applications in viticulture-a review. Aust. J. Grape Wine Res. 8 (1), 36–47. http://dx.doi.org/10.1111/j.1755-0238.2002.tb00209.x.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778. http://dx.doi.org/10.48550/arXiv.1512.03385.

Howard, J., Gugger, S., 2020. Fastai: A layered API for deep learning. Information 11 (2), 108. http://dx.doi.org/10.3390/info11020108.

ISP STIIMA, 2023. Vine trimming dataset. URL https://github.com/ispstiima/VineTrimmingDataset/.

Jafarbiglu, H., Pourreza, A., 2022. A comprehensive review of remote sensing platforms, sensors, and applications in nut crops. Comput. Electron. Agric. 197, 106844. http://dx.doi.org/10.1016/j.compag.2022.106844.

Kerkech, M., Hafiane, A., Canals, R., 2020. Vine disease detection in UAV multispectral images using optimized image registration and deep learning segmentation approach. Comput. Electron. Agric. 174, 105446. http://dx.doi.org/10.1016/j.compag.2020.105446.

King, P., Smart, R., McClellan, D., 2014. Within-vineyard variability in vine vegetative growth, yield, and fruit and wine composition of Cabernet Sauvignon in Hawke's Bay, New Zealand. Aust. J. Grape Wine Res. 20 (2), 234–246. http://dx.doi.org/10.1111/ajgw.12080.

Li, R., Zheng, S., Zhang, C., Duan, C., Su, J., Wang, L., Atkinson, P.M., 2021. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–13. http://dx.doi.org/10.1109/TGRS.2021.3093977.

Lin, Y., 2015. LiDAR: An important tool for next-generation phenotyping technology of high potential for plant phenomics? Comput. Electron. Agric. 119, 61–73. http://dx.doi.org/10.1016/j.compag.2015.10.011.

Liu, W., Li, Y., Liu, J., Jiang, J., 2021. Estimation of plant height and aboveground biomass of toona sinensis under drought stress using RGB-D imaging. Precis. Agric. 12, 1747. http://dx.doi.org/10.3390/f12121747.

Lu, Y., Young, S., 2020. A survey of public datasets for computer vision tasks in precision agriculture. Comput. Electron. Agric. 178, 105760. http://dx.doi.org/10.1016/j.compag.2020.105760.

Marani, R., Milella, A., Petitti, A., Reina, G., 2021. Deep neural networks for grape bunch segmentation in natural images from a consumer-grade camera. Precis. Agric. 22, 387–413. http://dx.doi.org/10.1007/s11119-020-09736-0.

Matese, A., Filippo Di Gennaro, S., 2015. Technology in precision viticulture: A state of the art review. Int. J. Wine Res. 69–81. http://dx.doi.org/10.2147/IJWR.S69405.

Matese, A., Toscano, P., Di Gennaro, S., Genesio, L., Vaccari, F., Primicerio, J., Belli, C., Zaldei, A., Bianconi, R., Gioli, B., 2015. Intercomparison of UAV, aircraft and satellite remote sensing platforms for precision viticulture. Remote Sens. 7 (3), 2971–2990. http://dx.doi.org/10.3390/rs70302971.

Microsoft, 2022. Azure kinect sensor SDK. URL https://learn.microsoft.com/en-us/azure/kinect-dk/sensor-sdk-download.

Milella, A., Marani, R., Petitti, A., Reina, G., 2019. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. Comput. Electron. Agric. 156, 293–306. http://dx.doi.org/10.1016/j.compag.2018.11.026.

Mohimont, L., Alin, F., Rondeau, M., Gaveau, N., Steffenel, L.A., 2022. Computer vision and deep learning for precision viticulture. Agronomy 12 (10), 2463. http://dx.doi.org/10.3390/agronomy12102463.

Moreno, H., Andújar, D., 2023. Proximal sensing for geometric characterization of vines: A review of the latest advances. Comput. Electron. Agric. 210, 107901. http://dx.doi.org/10.1016/j.compag.2023.107901.

Negassi, M., Wagner, D., Reiterer, A., 2022. Smart (sampling) augment: Optimal and efficient data augmentation for semantic segmentation. Algorithms 15 (5), 165. http://dx.doi.org/10.3390/a15050165.

Oliver, M.A., Bishop, T.F., Marchant, B.P., 2013. Precision Agriculture for Sustainability and Environmental Protection. Routledge Abingdon.

Palacios, F., Bueno, G., Salido, J., Diago, M.P., Hernández, I., Tardaguila, J., 2020. Automated grapevine flower detection and quantification method based on computer vision and deep learning from on-the-go imaging using a mobile sensing platform under field conditions. Comput. Electron. Agric. 178, 105796. http://dx.doi.org/10.1016/j.compag.2020.105796.

Pallottino, F., Antonucci, F., Costa, C., Bisaglia, C., Figorilli, S., Menesatti, P., 2019. Optoelectronic proximal sensing vehicle-mounted technologies in precision agriculture: A review. Comput. Electron. Agric. 162, 859–873. http://dx.doi.org/10.1016/j.compag.2019.05.034.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32, http://dx.doi.org/10.48550/arXiv.1912.01703.

Prechelt, L., 2002. Early stopping-but when? In: Neural Networks: Tricks of the Trade. Springer, pp. 55–69. http://dx.doi.org/10.1007/978-3-642-35289-8_5.

Qi, Y., Dong, X., Chen, P., Lee, K.-H., Lan, Y., Lu, X., Jia, R., Deng, J., Zhang, Y., 2021. Canopy volume extraction of Citrus reticulate Blanco cv. Shatangju trees using UAV image-based point cloud deep learning. Precis. Agric. 13, 3437. http://dx.doi.org/10.3390/rs13173437.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Roy, R.M., Ameer, P., 2021. Segmentation of leukocyte by semantic segmentation model: A deep learning approach. Biomed. Signal Process. Control 65, 102385. http://dx.doi.org/10.1016/j.bspc.2020.102385.

Santos, T., de Souza, L., dos Santos, A., Avila, S., 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Comput. Electron. Agric. 170, 105247. http://dx.doi.org/10.1016/j.compag.2020.105247.

Sishodia, R.P., Ray, R.L., Singh, S.K., 2020. Applications of remote sensing in precision agriculture: A review. Remote Sens. 12 (19), 3136. http://dx.doi.org/10.3390/rs12193136.

Stafford, J.V., 2000. Implementing precision agriculture in the 21st century. J. Agric. Eng. Res. 76 (3), 267–275. http://dx.doi.org/10.1006/jaer.2000.0577.

Vulpi, F., Marani, R., Petitti, A., Reina, G., A., M., 2022. An RGB-D multi-view perspective for autonomous agricultural robots. Comput. Electron. Agric. 202, 107419. http://dx.doi.org/10.1016/j.compag.2022.107419.

W3schools, 2023. Python Random shuffle Method URL https://www.w3schools.com/python/ref_random_shuffle.asp.

Weiss, K., Khoshgoftaar, T.M., Wang, D., 2016. A survey of transfer learning. J. Big data 3 (1), 1–40. http://dx.doi.org/10.1186/s40537-016-0043-6.

Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1492–1500. http://dx.doi.org/10.48550/arXiv.1611.05431.

Xu, Y., Tong, X., Stilla, U., 2021. Voxel-based representation of 3D point clouds: Methods, applications, and its potential use in the construction industry. Autom. Constr. 126, 103675. http://dx.doi.org/10.1016/j.autcon.2021.103675.