iVES OENO One vine & wine

DOI: https://doi.org/10.20870/oeno-one.2026.60.1.9628

OPEN ACCESS

VITICULTURE ORIGINAL RESEARCH ARTICLES

Grapevine-Seg: A grapevine segmentation method based on an improved YOLACT

Lingxin Bu[1], Jie Su[2], Qiangqiang Zhang[3], Qianwen Kou[1], Yun Chen[1], Jipeng Wang[1], Xingrun Tang[1], Xingjia Li[1], Xingtao Wu[1], Teng Zhang[1,*], and Li Zhao[1]

[1] College of Mechatronic Engineering, North Minzu University, Yinchuan, Ningxia, 750021, China
[2] Ningxia Hongyuan Great Wall Machine Tool Co., Ltd., Yinchuan, Ningxia, 750021, China
[3] School of Mechanical Engineering, Ningxia University, Yinchuan, Ningxia, 750021, China

Article number: 9628

*correspondence:
zhangteng1893@163.com

Associate editor:
Bruno Tisseyre

Received:
21 September 2025

Accepted:
2 February 2026

Published:
12 March 2026

This article is published under the **Creative Commons licence** (CC BY 4.0).

Use of all or part of the content of this article must mention the authors, the year of publication, the title, the name of the journal, the volume, the pages and the DOI in compliance with the information given above.

ABSTRACT

Real-time segmentation of grapevine cordons and shoots remains a critical bottleneck for automated viticultural management, particularly in resource-constrained field environments where computational efficiency is paramount. This study established a lightweight segmentation model based on an improved YOLACT (You Only Look At Coefficient) framework, optimised for real-time grapevine segmentation on embedded systems. In place of the original ResNet backbone is Ghost-Net, which uses Ghost modules to generate redundant feature maps and produce richer feature representations, thus reducing the computational load of convolutions and improving small-object detection and segmentation. Concurrently, an EMAttention (Efficient Multi-scale Attention) module is embedded at the skip connections between the Feature Pyramid Network (FPN) and the Mask Head. This module concatenates the FPN features of multiple scales with Mask Head features, applies spatial convolution to generate spatial attention maps that bolster object-region features, and adaptively fuses multi-scale features with learned weights, thus improving segmentation performance across differently sized objects. The modifications, focusing on lightweight design and inference speed, yield a model that balances accuracy and efficiency more suitably for embedded viticultural systems than existing benchmarks. Under identical experimental conditions, the improved model outperformed mainstream segmentation frameworks. The average detection accuracies for grapevine and lateral shoot test samples reached 69.46 % and 67.66 %, respectively. These results demonstrate the potential of this approach for enabling real-time, field-deployable grapevine segmentation in precision viticulture applications. The annotated dataset (Grapevine-Seg) is publicly available at https://zenodo.org/records/18218165.

KEYWORDS: deep learning, convolutional network, vineyard, precision viticulture, grapevine pruning

INTRODUCTION

Grapevines, as economically significant crops, require precise management to optimise yield and improve quality. Grapevine segmentation serves as a core technology enabling precision operations and has been a research focus within agricultural intelligence for several decades. Early pioneering work on vision-guided grapevine pruning can be traced back to Mercurio *et al.* (1989), who developed a block-type robotic pruner with machine vision capabilities. Subsequently, McFarlane *et al.* (1997) explored image analysis techniques for pruning long-wood grape vines, and Gao and Lu (2006) further investigated image processing methods for autonomous grapevine pruning. These early studies highlighted the inherent complexity of extracting meaningful features from vine images and laid the foundation for subsequent advances. With the emergence of deep learning, this field has attracted renewed attention and has recently become an active research area (Botterill *et al.*, 2017). Traditional viticulture relies heavily on manual labour for essential tasks like pruning and flower thinning, which is not only inefficient and costly but also unsustainable given large-scale production demands (Bochtis *et al.*, 2014). Thus, a growing need exists to advance automation and intelligence in viticultural management (Bu *et al.*, 2025; Íñiguez *et al.*, 2025).

In viticultural management research, Karkee *et al.* (2023) applied robotic precision pruning to planar tree canopies in orchards and vineyards. Their approach optimises yield through load management techniques such as pruning and thinning, thereby increasing productivity and resource utilisation. Furthermore, Guadagna *et al.* (2023) used the Faster Region-based Convolutional Neural Network (Faster R-CNN) model to detect visible inter-mediate complex buds with a detection rate of 0.97 and the most prevalent coplanar simple buds at 74 %. Mask Region-based Convolutional Neural Network (Mask R-CNN) experiments indicated optimal node segmentation and a shoot-thinned recall rate of 0.85, exceeding the control group. Similarly, Majeed *et al.* (2019) developed a Faster R-CNN–based method to detect visible stem segments using transfer learning on pretrained networks. The Residual Network 18 (ResNet-18)-based model performed the best, achieving an F1 score of 0.55 and a mean average precision of 45.1 %. Majeed *et al.* (2020a) used deep learning networks to determine the grapevine's main stem contour from colour camera imagery, applying SegNet and FCN (Fully Convolutional Networks) segmentation techniques. The model accurately traced the main stem trajectory even when concealed by foliage. In a different study, Majeed *et al.* (2020b) also detected visible segments of the trunk and main stem using Faster R-CNN with a ResNet-18 backbone, applied non-maximum suppression to refine detections, and fitted a 6th-degree polynomial to centroids of detected segments to estimate the main stem trajectory. They found that for vines two to four weeks post-budburst, trajectory estimation yielded correlation coefficients of 0.993, 0.991, and 0.987. Moreover, Marset *et al.* (2021) implemented a fully convolutional network based on MobileNet for bud segmentation. They performed pixel-level classification, followed by post-processing, to establish bud correspondences and centroid localisation. The best FCN-MN (Fully Convolutional Networks MobileNet) model achieved an F1 score of 88.6 %, indicating high segmentation accuracy. Similarly, Moreno and Andújar (2023) reviewed the application of proximal sensing technologies for geometric characterisation of grapevines, concluding that while LiDAR offers high precision but at an elevated cost, ultrasonic sensors are low-cost and low-resolution, and that depth cameras achieve a balance between cost and accuracy. Furthermore, Casado-García *et al.* (2022) compared multiple depth architectures and employed three semi-supervised techniques, including pseudo-labelling, leveraging unlabelled data. The semi-supervised approach increased the average accuracy by 5.62 %–6.01 %. Gentilhomme *et al.* (2023) used the ViNet deep learning approach, employing a stacked hourglass network to detect nodes, classify branch types, and infer spatial relationships. A shortest-path weighted graph algorithm was then employed to optimally extract connections between nodes, yielding a node precision of 95 % and a recall of 90 %. Dong *et al.* (2016) adopted a Mask R-CNN segmentation method to compile a dataset of key grapevine structures and conducted a comparative analysis. Their Mask R-CNN model, featuring a Residual Network 101 (ResNet-101) + feature pyramid network (FPN) backbone, achieved precision, recall, and mean average precision values of 85.04 %, 82.03 %, and 85.40 %, respectively, significantly outperforming comparative models. Recent studies have demonstrated that combining automatic annotation pipelines with transformer-based segmentation frameworks can significantly improve structural delineation and growth analysis in field crops (Rana *et al.*, 2024). Such integrated methods using YOLO-based detectors and SAM (Segment Anything Model) architectures have proven effective in capturing fine structural boundaries across complex agricultural scenes, and similar workflows can be adapted for grapevine canopy and cordon segmentation. More recently, Fernandes *et al.* (2025) demonstrated the potential of merging 2D segmentation with 3D point clouds for pruning point generation, representing a promising direction for integrating multi-modal data in viticultural applications.

Although existing models have made progress in accuracy and environmental robustness, the dynamic nature of grapevine growth—such as structural variations across phenological stages—and complex environmental interference in field settings (*e.g.*, weeds and uneven lighting) continue to limit the practical application of segmentation techniques. To address these challenges, this study introduces an improved YOLACT-based framework by innovatively integrating GhostNet and EMAttention mechanisms. In contrast to YOLACT++, which primarily focuses on backbone refinement, or recent approaches such as ViNet that employ more complex architectures, our lightweight solution leverages Ghost modules to generate redundant feature maps, thereby enhancing small-object feature representation while reducing computational cost.

Simultaneously, the EMAttention module enables adaptive fusion of multi-scale features, improving segmentation accuracy across varying growth stages and complex field environments without compromising real-time performance. Further investigation into efficient segmentation algorithms remains essential to advance intelligent viticultural systems.

This work is practical rather than conceptual. We present an incremental yet effective improvement to the YOLACT architecture, specifically optimised for the challenge of real-time grapevine segmentation in resource-constrained field environments. Specifically, the scientific objectives of this study are: (1) to develop a lightweight segmentation model capable of real-time grapevine cordon and shoot detection suitable for deployment on resource-constrained embedded systems; (2) to evaluate whether the integration of GhostNet backbone and EMAttention mechanism can effectively improve segmentation accuracy for small and variable-sized grapevine structures under complex field conditions; and (3) to establish a publicly available annotated dataset (Grapevine-Seg) that can serve as a benchmark for future research in automated viticultural management.

MATERIALS AND METHODS

1. Image acquisition and preprocessing

Currently, publicly available datasets for grapevine instance segmentation are limited. Gentilhomme *et al.* (2023) collected 1,513 images of grape plants using smartphones or digital cameras, annotating structural elements such as the trunk, cane, shoots, and nodes for structural extraction tasks. In the Helan Mountains west slope region of Ningxia, vineyards primarily leverage a "sloping trunk horizontal cordon" training system, which aids in facilitating vine burial and emergence while mitigating frost injury and desiccation, as shown in Figure 1.

Images were captured in the vineyard of Ningxia Mutong Winery Co., Ltd., at the coordinates 38.61° N 106.13° E. The grape cultivar imaged was Cabernet-Sauvignon, cultivated for 7–8 years. Image capture was conducted using a Huawei Nova 11 smartphone. Each image measured 1,920 × 1,080 pixels, and data collection spanned from 09:00 to 17:00 on October 28, 2023. Then, the original images were annotated using LabelMe. Structural components, including the cordon and shoots, were delineated to produce JSON annotation files, which were then converted into segmentation mask images. The final dataset comprised 2,091 images, which were divided into a training set of 1,672 images and a validation set of 419 images. The dataset is publicly available on https://zenodo.org/records/18218165 for academic research and can thus be used for benchmarking purposes.

2. Improved YOLACT model

YOLACT is the first single-stage algorithm that enables real-time instance segmentation by predicting prototype masks and mask coefficients, which are linearly combined to synthesise the final segmentation masks (Bolya *et al.*, 2019). Its key components are the backbone network, feature pyramid network (FPN), Detection Head, and Mask Head. However, its architecture presents several limitations that hinder its effectiveness in practical applications like automated viticulture: (1) The commonly used ResNet-50 backbone suffers from low efficiency, with 25.6 M parameters and 3.8 G FLOPs, creating significant computational redundancy and hindering deployment on edge devices. (2) The feature pyramid network (FPN) primarily performs multi-scale fusion but lacks sophisticated attention mechanisms to model channel-wise semantics and spatial instance specificity, limiting its feature representation power. (3) The simplified prototype mask generation, a design choice for speed, results in insufficient capture of fine-grained features, leading to a suboptimal speed-accuracy trade-off. To address the identified limitations, two key enhancements are introduced into the YOLACT framework: 1) GhostNet for computational efficiency:

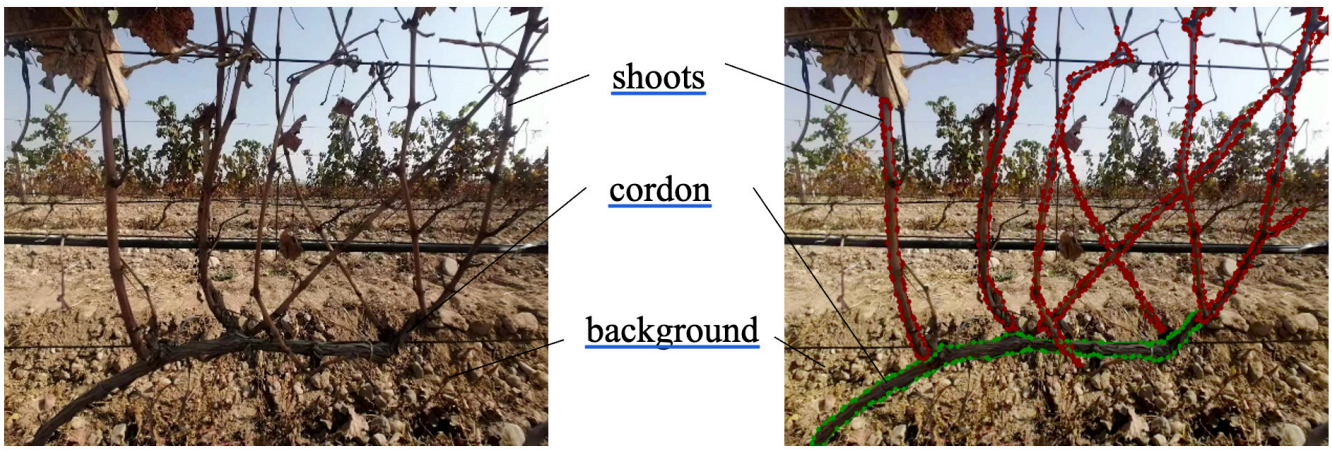


FIGURE 1. Sample image from the grapevine dataset comprising: (a) the original RGB image of the grapevine plant and (b) the annotated mask with structural labels such as shoots and cordon.

the ResNet backbone was replaced with GhostNet, which employs Ghost modules (Han *et al.*, 2020). These modules use a combination of base convolutions and cost-effective linear transformations to generate feature maps with reduced redundancy. This architecture decreases the parameter count to 5.2 M (1/5 of ResNet-50) and FLOPs to 0.4 G (1/9 of ResNet-50), while maintaining competitive feature representation. The superior semantic consistency of its multi-scale features facilitates faster inference, making the model suitable for low-power devices. 2) EMAttention for feature enhancement: the dual branch structure of "ECA channel attention and spatial attention" (Ouyang *et al.*, 2023). The ECA (Efficient Channel Attention) channel attention branch employs adaptive 1D convolution to efficiently capture cross-channel interactions with a computational cost only 1/8 of the SE (Squeeze-and-Excitation) module. The spatial attention branch uses depth-wise convolution to generate pixel-wise weight maps, effectively suppressing irrelevant background clutter. This module is designed to improve mask accuracy in challenging scenarios, such as occlusions, with minimal impact on model size and inference speed.

2.1. Lightweight backbone

The original ResNet backbone is replaced with GhostNet, incorporating Ghost modules to produce redundant feature maps while reducing convolutional computation. Specifically, GhostNet employs a two-pronged approach: a 1×1 convolution generates a subset of intrinsic feature maps, which are subsequently expanded via cheap linear operations, such as depth-wise separable convolutions, to increase channel dimensions significantly, and with markedly fewer parameters. The Φ_i is the identity mapping for preserving the intrinsic feature maps, as shown in Figure 2 (Han *et al.*, 2020).

2.2. Multi-scale attention enhancement

An EMAttention module is embedded at the skip connection between the Mask Head and the FPN, as depicted in Figure 3 (Ouyang *et al.*, 2023). This module dynamically calibrates channel weights during multi-scale feature fusion through group shuffle and cross-dimensional interactions, thus refining target edges and fine-grained details.

Given an input feature map F with dimensions $C \times H \times W$, where C denotes the number of channels, and H and W represent the height and width, respectively, G means the divided groups, the attention mechanism functions as follows:

(1) channel attention: channel descriptors are produced using global average pooling and global max pooling. These descriptors are passed through fully connected layers, and channel-wise weights are found using a sigmoid activation function.

(2) spatial attention: pooling is conducted across the channel dimension to produce spatial feature maps, which are then processed by convolutional layers to generate spatial weight maps, followed by sigmoid activation. After each training step, the feature map weights are updated using Exponential Moving Average (EMA), computed as shown in Equation 1:

$$W_{new} = \alpha W_{old} + (1 - \alpha)W \qquad (1)$$

where W_{new} is the updated attention weight, W_{old} signifies the attention weight at the previous moment, W represents the attention weight at the current moment, and α is the decay factor, typically ranging from 0.9 to 0.99. The attention weights are applied to the feature map F as shown in Equation 2. The output F' is the weighted feature map F, which is then passed to subsequent network layers for further processing.

$$F' = FI_{CW}I_{SW} \qquad (2)$$

where I_{CW} and I_{SW} are the channel weight and spatial weight, respectively.

The EMA module employs parallel substructures and a shared 1×1 convolutional branch within the CA (Coordinate Attention) module to mitigate extensive sequential processing and excessive network depth. For aggregating multi-scale spatial structural information, EMA strategically places a 3×3 convolutional kernel in parallel with the 1×1 branch, effectively establishing both short-range and long-range dependencies to enhance performance.

The improved model is visualised in Figure 4. The model takes an image of size 550 × 550 as input. First, the GhostNet backbone extracts multi-scale base feature maps (denoted as C1–C5) at different down-sampling stages.

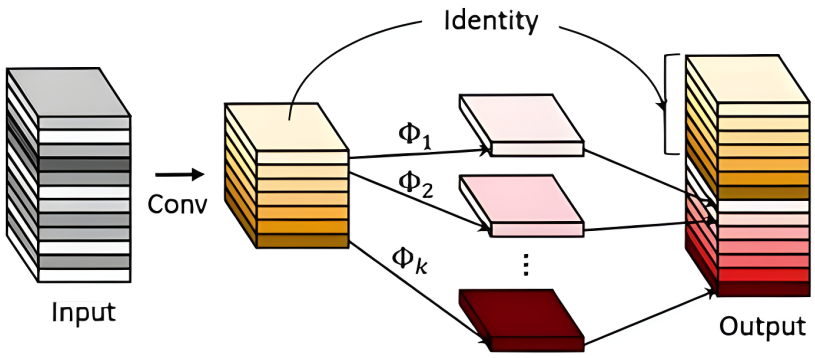


FIGURE 2. Schematic diagram of the GhostNet architecture. (The Φ_i is the identity mapping for preserving the intrinsic feature maps.)

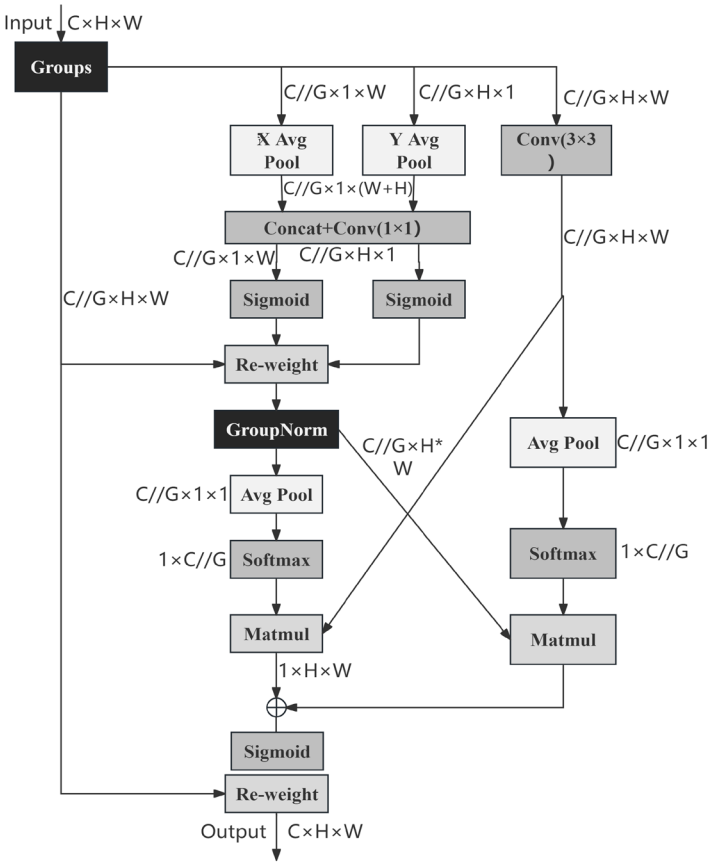


FIGURE 3. Schematic diagram of the EMAttention module structure.

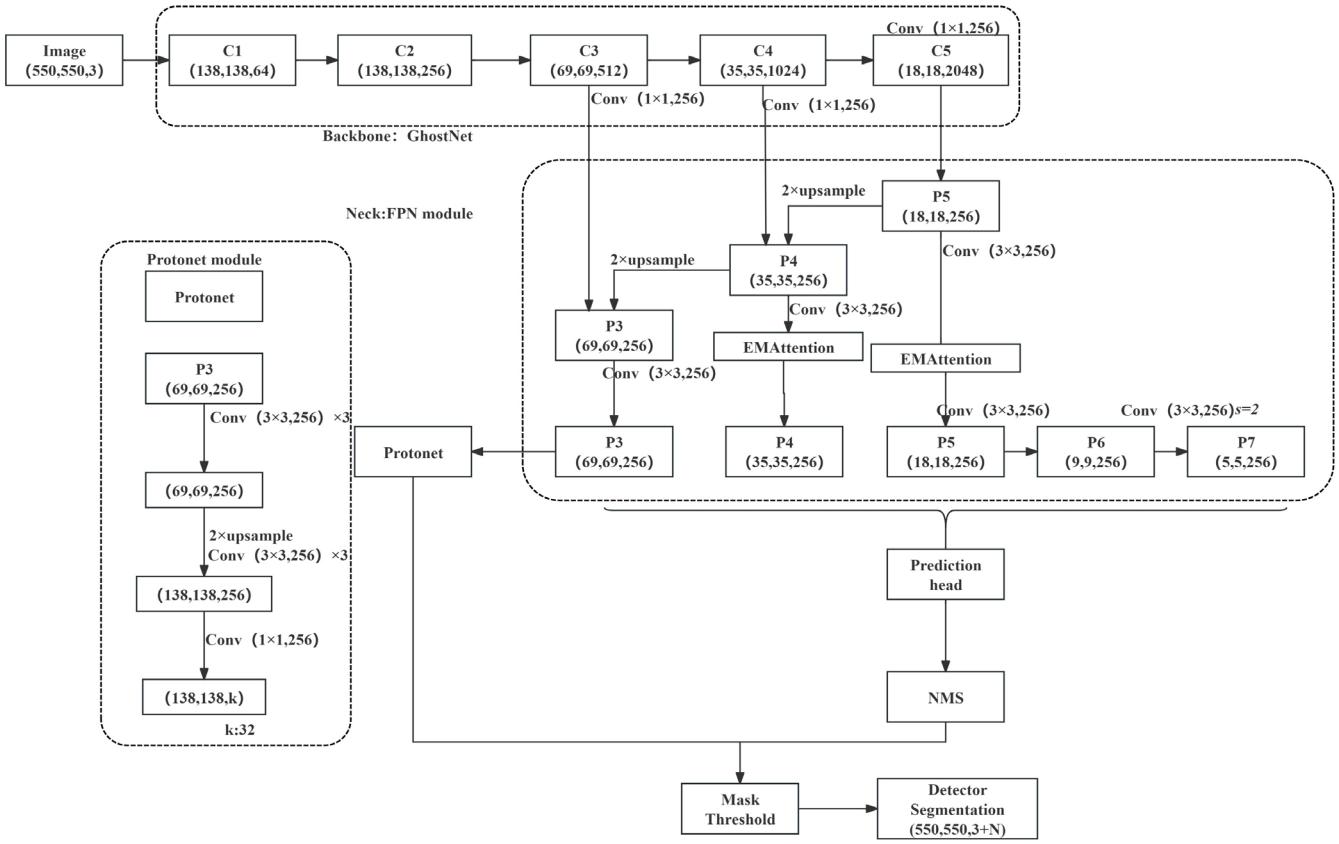


FIGURE 4. Network topology of the improved model.

Subsequently, a feature pyramid network (FPN) performs up-sampling, fusion, and enhancement on C3–C5 to generate integrated feature maps (denoted as P3–P7) that combine rich semantic information with fine spatial details. The model accomplishes the task through two parallel branches: the detection head produces classification scores and bounding box regressions based on P3–P7, while simultaneously predicting a set of 32 mask coefficients for each instance; the prototype mask head (Protonet) generates 32 globally shared prototype masks from P3–P7. The final instance mask is obtained via a linear combination of the prototype masks and the corresponding instance-specific mask coefficients. The results are then post-processed through non-maximum suppression (NMS), mask cropping, and up-sampling to yield pixel-wise instance segmentation masks aligned with the input resolution. The Ghost module generates additional feature maps through inexpensive operations, thereby markedly reducing the model's computational load. The Ghost module produces rich feature representations while keeping computational costs low, thus improving the performance of small-object detection and segmentation. Moreover, GhostNet's lightweight nature makes it more suitable for mobile devices and edge computing scenarios, thereby improving the model inference speed and facilitating practical applications.

The EMAttention module is introduced at the skip connection between the Mask Head and the FPN, concatenating multi-scale FPN features with Mask Head features. Spatial convolutions generate spatial attention maps to refine target region features. Global pooling and multilayer perceptrons generate channel attention maps to highlight important channels. Furthermore, adaptive weighted fusion of multiscale features improves segmentation capability for differently sized objects.

3. Model training

The experimental hardware platform was a Dell workstation equipped with an Intel Xeon E5-1620 processor, 32 GB of RAM, and an Nvidia GeForce RTX 2080 Ti graphics card. The operating system was Ubuntu 18.04, and the deep learning framework was PyTorch 1.2 with Python 3.6. The specific model training parameters are detailed in Table 1.

4. Evaluation metrics

The evaluation metrics used to assess the model's performance were precision *(P)*, recall *(R)*, average precision *(AP)*, mean average precision *(mAP)*, parameter count, and detection frame rate. Precision is the proportion of true-positive detections among all predicted positive instances, serving as an indicator of the model's accuracy in identifying relevant objects. Recall represents the proportion of true-positive detections among all actual positive instances, reflecting the model's ability to detect all relevant objects. *AP* reflects the model's performance across different recall levels, computed by integrating the precision–recall curve *(P-R)*. *mAP* is the mean of *AP* values across all object classes, offering an overall assessment of the model's detection capabilities. Additionally, the parameter count and detection frame rate reflect the model's computational efficiency and real-time processing capability. These metrics collectively provide a holistic view of the model's effectiveness in object detection tasks. The evaluation metrics were calculated according to Equations 3–6.

$$P = \frac{T_P}{T_P + F_P} \tag{3}$$

$$R = \frac{T_P}{T_P + F_N} \tag{4}$$

$$AP = \int_0^1 P(R)\ dR \tag{5}$$

$$mAP = \frac{\sum_{i=1}^{l} AP_i}{l} \tag{6}$$

In the formula, T_P denotes the number of true-positive detections; F_P represents the number of false-positive detections; F_N signifies the number of false-negative detections; l refers to the number of detection categories; and *AP* is the area under the *P*-*R* curve, where *R* is plotted on the *x*-axis, and *P* is shown on the *y*-axis. *AP* reflects a comprehensive measure of model performance.

For automated pruning and yield estimation, minimising false negatives (*i.e.*, maximising recall) is often more critical. In dormant pruning, missing a shoot that requires removal *(F_N)* means it will be left on the vine, potentially leading to incorrect canopy structure, wasted nutrients, and compromised yield/quality in the following season. Similarly, for yield estimation, failing to count a shoot *(F_N)* directly leads to an underestimation of the potential yield, preventing accurate management decisions. Therefore, a high recall ensures that most management targets are captured. For generating precise maps or for robot navigation, minimising false positives (*i.e.*, maximising precision) can be more important. If the model mistakes shadows or weeds for shoots *(F_P)*, a canopy map generated from this information would be noisy and misleading for management. For a physical robot, attempting to cut a non-existent "shoot" is not only an inefficient action but also poses a risk of mechanical failure and energy waste.

TABLE 1. Model training parameters.

Parameter category	Parameter setting
Initial learning rate	2e-3
Weight decay rate	5e-4
Optimizer	SGD
Number of epochs	200

In this research, the average precision *(AP)* was chosen as a core metric precisely because it integrates both precision and recall across multiple confidence thresholds (*i.e.*, the area under the precision-recall curve). A high *mAP* score indicates that the model achieves a good overall balance between "finding all relevant objects (high recall)" and "ensuring the detected objects are correct (high precision)", which is most desirable for a versatile automation system expected to perform multiple tasks.

RESULTS AND DISCUSSION

To evaluate the impact of different attention mechanisms on the YOLACT model, the study incorporated Squeeze-and-Excitation (SE) (Hu *et al.*, 2018), Coordinate Attention (CA) (Hou *et al.*, 2021), Efficient Channel Attention (ECA) (Wang *et al.*, 2020), and EMAttention modules into the YOLACT architecture. The comparative experimental results are given in Table 2. As shown, the introduction of different attention modules led to a certain degree of improvement in the segmentation performance of the original model. Of these models, the EMAttention module exhibited the greatest improvement, with *P*, *R*, and *mAP* increasing to 74.5 %, 64.2 %, and 67.98 %, respectively.

To evaluate the impact of different lightweight backbone networks on the segmentation and detection performance of the model, this study replaced the original backbone with MobileNetV2 (Sandler *et al.*, 2018), MobileNetV3 (Howard *et al.*, 2019), ShuffleNetV2 (Ma *et al.*, 2018), and GhostNet. The comparative experimental results are presented in Table 3. As shown, the GhostNet backbone achieves the best performance across all metrics, with *P*, *R*, *mAP*, and parameter count improving by 1.7, 1.9, and 1.04 percentage points, respectively, compared with the original model. Additionally, the model's parameter count was reduced to 35.64 MB, or a decrease of 15.78 %.

Figure 5 illustrates the loss curves during training for four models: YOLACT, YOLACT + GhostNet, YOLACT + EMAttention, and YOLACT + GhostNet + EMAttention. The curves reveal that the YOLACT + GhostNet + EMAttention model converged more rapidly and achieved the lowest loss on the test set compared with the other models. Table 4 lists the performance metrics of these models, demonstrating that the proposed modifications improved the segmentation performance. In order to avoid the randomness of the results caused by a single training, each training is repeated three times, and the average and standard deviation of the three training results are calculated. Specifically, the improved model achieved a mask *mAP* of 67.83 ± 0.61 %, signifying a 4.00 percentage point increase over the original model. The detection *mAP* was 68.57 ± 0.74 %, or a 3.75 percentage point improvement. These results underscore the effectiveness of integrating GhostNet and EMAttention modules in augmenting the segmentation capabilities for grapevine and shoot instances. Figure 6 provides a visual comparison of segmentation results for grapevine and shoots across the four models. The enhanced model demonstrates superior segmentation accuracy, particularly in delineating fine structures and small targets. Furthermore, the optimised model operated at a detection speed of 46.72 frames per second (FPS), representing a 17.59 % increase over the original model. This improvement aligns with the goal of achieving both higher segmentation quality and a faster detection speed.

To evaluate the performance of the proposed model on grape images, we conducted comparative experiments against several mainstream segmentation models to assess its effectiveness and generalisation capability. Specifically, the study selected Mask R-CNN, YOLACT++, BlendMask, and SOLOv2 for comparison. As shown in Table 5, the proposed model demonstrated significant advantages in both mask *mAP* and bounding box, achieving 67.83 ± 0.61 % and 68.57 ± 0.74 %, respectively. The detection accuracy was 75.7 ± 0.75 %, surpassing that of the other models. Although the *R* of the improved model was 2.15 percentage points lower than that of BlendMask, it outperformed Mask R-CNN, YOLACT++, and SOLOv2 by 3.45, 2.42, and 1.37 percentage points, respectively. Furthermore, the bounding box *mAP* of the improved model exceeded that of Mask R-CNN, YOLACT++, BlendMask, and SOLOv2 by 4.88, 3.89, 3.96, and 3.29 percentage points, respectively.

TABLE 2. Comparative experiment of integrating different attention modules into the YOLACT network.

Attention modules	*P*/%	*R*/%	*mAP*/%
YOLACT	72.1	63.3	64.83
SE	72.9	64.8	65.32
CA	73.3	63.9	65.67
ECA	73.2	64.0	66.03
EMAttention	74.5	64.2	67.98

TABLE 3. Comparative experiment of the YOLACT network with different backbone networks.

Backbone networks	*P*/%	*R*/%	*mAP*/%	Model parameter/MB
MobileNetV2	72.5	63.9	64.88	36.75
MobileNetV3	73.0	64.8	65.27	38.66
ShuffleNetV2	73.3	64.6	65.31	39.06
GhostNet	73.8	65.2	65.87	35.64

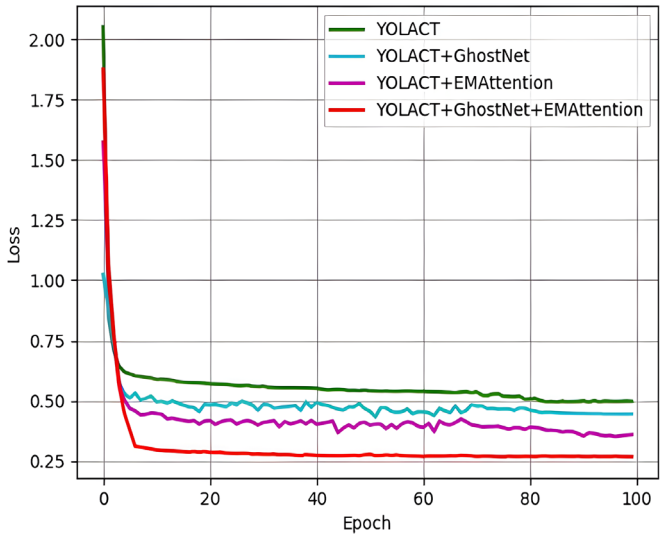


FIGURE 5. Loss curves during the training process of four models.

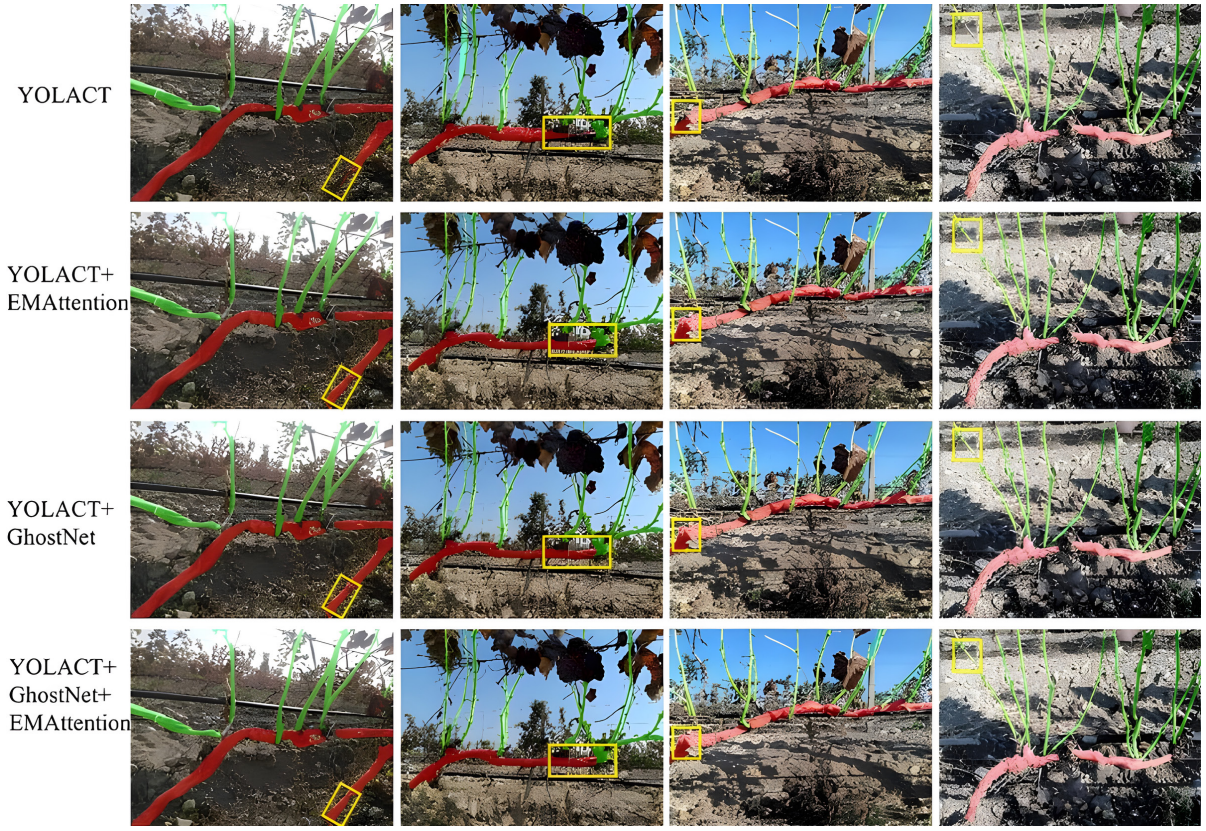


FIGURE 6. Comparison of grapevine segmentation results across four models.

Note: the yellow boxes indicate areas where the original model failed to accurately segment the grapevine and side branches, whereas the improved model achieved precise segmentation.

TABLE 4. Ablation test results.

Model	*P*/%	*R*/%	Bounding box *mAP*/%	Bounding box *AP75*/%	Mask *mAP*/%	Mask *AP75*/%	Detection speed/FPS
YOLACT	72.13 ± 1.35	63.20 ± 0.36	64.82 ± 0.06	75.51 ± 0.03	63.83 ± 0.25	71.04 ± 0.20	39.73
YOLACT + GhostNet	73.43 ± 0.72	64.17 ± 1.17	65.88 ± 0.62	77.73 ± 0.11	65.25 ± 0.30	72.95 ± 1.69	47.85
YOLACT + EMAttention	73.30 ± 1.82	64.73 ± 0.68	67.31 ± 0.94	78.27 ± 0.09	63.78 ± 0.35	70.82 ± 0.69	38.77
YOLACT + GhostNet + EMAttention	75.70 ± 0.75	66.5 ± 0.75	68.57 ± 0.74	79.16 ± 0.88	67.83 ± 0.61	73.60 ± 0.83	46.72

TABLE 5. Comparative performance of five models in the grapevine image.

Model	*P*/%	*R*/%	Bounding box *mAP*/%	Bounding box *AP75*/%	Mask *mAP*/%	Mask *AP75*/%	Detection speed/FPS
Mask R-CNN	68.45 ± 0.85	63.05 ± 0.13	63.69 ± 0.19	74.64 ± 0.98	62.94 ± 1.17	69.47 ± 1.40	23.73
YOLACT++	73.52 ± 0.38	64.08 ± 0.25	64.68 ± 0.64	75.62 ± 0.45	63.62 ± 1.00	70.90 ± 0.59	39.73
BlendMask	69.35 ± 0.38	68.65 ± 0.49	64.61 ± 0.29	76.85 ± 0.55	60.95 ± 0.35	65.99 ± 2.51	38.09
SOLOv2	74.77 ± 0.40	65.13 ± 0.25	65.28 ± 1.65	77.53 ± 0.19	65.51 ± 0.85	70.47 ± 0.27	43.05
Ours	75.70 ± 0.75	66.50 ± 0.75	68.57 ± 0.74	79.16 ± 0.88	67.83 ± 0.61	73.60 ± 0.83	46.72

In terms of detection speed, the model operated at 46.72 FPS, representing improvements of 96.88 %, 17.59 %, 22.66 %, and 8.52 % over Mask R-CNN, YOLACT++, BlendMask, and SOLOv2, respectively. Thus, the proposed model not only improved segmentation performance but also achieved higher detection speed, indicating its suitability for real-time applications in grapevine image analysis. SOLOv2 model is superior to other models in terms of *P*, *mAP* and detection speed. The YOLACT++ model is improved by introducing the GhostNet and EMAttention module. Therefore, the improved model has a better detection effect.

To validate the performance of the improved lightweight model proposed in this study on embedded devices, a Jetson Xavier NX developer kit (memory: 8 GB; GPU: 384-core NVIDIA Volta GPU with 48 Tensor Cores; CPU: 6-core NVIDIA Carmel ARM v8.2) running Ubuntu 22.04 was employed. The segmentation and detection speed of the improved model reached 2.34 FPS.

As shown in Table 6, the experimental results show that the proposed model achieves values of 69.09 ± 0.89 % for grapevine detection and 68.05 ± 0.72 % for shoot detection. These results represent improvements of 6.03, 4.62, 4.56, and 4.33 percentage points over Mask R-CNN, YOLACT++, BlendMask, and SOLOv2, respectively, in grapevine detection. For side branch detection, the improvements were 3.72, 3.17, 3.36, and 2.25 percentage points over the same models.

While the performance gains in detection accuracy and inference speed presented in this study may appear modest in absolute terms, their practical implications for vineyard operations are significant. SOLOv2, while effective for instance segmentation, demands considerable computational resources that can hinder deployment on embedded systems commonly used in agricultural robotics. Our improved YOLACT model achieves a better balance between accuracy and computational efficiency, making it more amenable for real-time applications under typical field conditions. In practice, real-time segmentation in vineyards requires not only high *mAP* but also sustained high inference speed on mid-range hardware to achieve responsive robotic control. Although SOLOv2 can achieve real-time performance on high-end GPUs, its deployment on cost-effective, power-efficient platforms—often necessary for field robots—remains challenging. The proposed model strikes a balance between competitive segmentation accuracy (~69 % *mAP*) and a high inference speed of 46.72 FPS, thereby facilitating its integration into automated vineyard systems. But the practical applicability of the model is currently constrained by the characteristics of the Grapevine-Seg dataset. The annotations were manually completed and, consequently, reflect specific grapevine varieties, trellising systems, and environmental conditions. An inherent class imbalance exists within the dataset, where shoot instances significantly outnumber cordon instances. Furthermore, potential geographic and seasonal biases persist. While data augmentation techniques were employed to mitigate these issues, they may still impact the model's generalisation capability. The public release of this dataset represents an initial step to address this limitation by encouraging the research community to incorporate more diverse data.

Furthermore, the ultimate objective of real-time grapevine segmentation extends beyond mere detection; it serves as a critical enabler for precision agriculture tasks such as automated pruning, yield estimation, and canopy management. 1) Automated pruning: the instance masks of shoots and the cordon generated by the model enable precise localisation of their junction points. This provides crucial spatial coordinates for a robotic cutting tool, allowing it to plan optimal cutting paths to selectively remove unwanted shoots while preserving healthy fruiting wood. 2) Shoot counting: the automatic count of segmented shoot instances facilitates the estimation of shoot density per unit length of cordon. This metric serves as a fundamental input for constructing accurate yield prediction models and for informing crop-load management decisions, such as cluster thinning. 3) Canopy management: the segmentation masks allow for the assessment of shoot distribution uniformity. By identifying overcrowded zones, the system can guide targeted shoot thinning operations. This optimises canopy light exposure and air circulation, thereby enhancing final fruit quality. Segmentation is, therefore, a prerequisite for the decision-making process of determining "which cane to cut" and "where to make the cut". The proposed method's improved speed-accuracy trade-off ensures that segmentation outputs can be processed within the tight latency constraints of closed-loop robotic control, thereby supporting continuous and adaptive operation in dynamic vineyard environments.

TABLE 6. Comparative experiment on the average precision of grapevines and shoots.

Model	Grapevines/%	Shoots/%
Mask R-CNN	63.06 ± 1.02	64.33 ± 0.78
YOLACT++	64.47 ± 1.93	64.88 ± 0.83
BlendMask	64.53 ± 1.72	64.69 ± 1.64
SOLOv2	64.76 ± 1.44	65.80 ± 2.08
Ours	69.09 ± 0.89	68.05 ± 0.72

CONCLUSION

In automated viticultural management, precise segmentation of vines and side branches is crucial for tasks such as pruning. To achieve rapid and accurate component segmentation, this study introduced the EMAttention and GhostNet modules into the YOLACT framework, resulting in a lightweight model that increases the detection speed and segmentation accuracy. Comparative experimental analyses yield the following conclusions:

1. The improved model achieved a bounding box *mAP* of 68.57 ± 0.74 %, mask *mAP* of 67.83 ± 0.61 %, and a detection speed of 46.72 FPS, representing increases of 3.75, 4.00 percentage points, and 17.59 %, respectively, over the original model. The average precision for grapevines and shoots was 69.09 ± 0.89 % and 68.05 ± 0.72 %, respectively.

2. The superior performance-efficiency balance achieved by our improved YOLACT-based model. When evaluated under consistent conditions, the proposed method exhibits a compelling advantage over leading segmentation approaches—including Mask R-CNN, YOLACT++, BlendMask, and SOLOv2—by simultaneously elevating detection accuracy (as reflected in bounding box *mAP*), enhancing mask quality, and reducing computational complexity. Notably, the model also delivers the fastest inference speed among all compared frameworks, thereby reinforcing its practical viability for real-time agricultural applications. Although BlendMask attained a marginally higher recall, the comprehensive gains across multiple metrics affirm that our approach offers a more effective and deployable solution for grapevine instance segmentation in resource-conscious field environments.

It should be noted that this study has certain limitations regarding the training and testing datasets. All images were collected from a single grape variety (Cabernet-Sauvignon), relatively young vines (7–8 years old), and a uniform training system (trellised Cordon de Royat). These constraints may affect the generalisability of the model to other viticultural contexts, as older vines typically exhibit greater structural complexity and different varieties may present distinct morphological characteristics.

Regarding practical applicability, the segmentation outputs generated by our model can be directly integrated into robotic systems for tasks such as automated pruning or thinning. For example, the instance masks corresponding to individual cane structures can be used by a path planning module to guide a robotic manipulator in making precise cuts during dormant pruning. Similarly, during canopy management, the segmented foliage regions can inform selective thinning operations to optimise sunlight exposure and air circulation. While this study focused on algorithm development, future work will include field validation via integration with a robotic platform equipped with a real-time perception-control loop. Prior to embedded deployment, it would be scientifically prudent to validate the proposed approach across diverse wine-growing regions worldwide, encompassing a wider range of grape varieties, vine ages, and training systems. Specifically, we plan to implement the proposed model on a pruning robot to evaluate its performance in operational scenarios, measuring task completion rates and robustness under varying field conditions.

ACKNOWLEDGEMENTS

The work in this paper was supported by the Natural Science Foundation of Ningxia (2025AAC030073, 2023AAC03302), Key Research and Development Project of Ningxia Hui Autonomous Region (2024BEH04137), and North Minzu University (2021KYQD31).

REFERENCES

Bochtis, D., Sørensen, C., & Busato, P. (2014). Advances in agricultural machinery management: A review. *Biosystems Engineering*, *126*, 69-81. https://doi.org/10.1016/j.biosystemseng.2014.07.012

Bolya, D., Zhou, C., Xiao, F., & Lee, Y. (2019). *YOLACT: Real-time instance segmentation*. Paper presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2019.00925

Botterill, T., Paulin, S., Green, R., Williams, S., Lin, J., Saxton, V., Mills, S., Chen, X., & Corbett-Davies, S. (2017). A robot system for pruning grape vines. *Journal of Field Robotics*, *34*(6), 1100-1122. https://doi.org/10.1002/rob.21680

Bu, L., Zhang, Q., Kou, Q., Chen, Y., Li, X., Wu, X., Zhang, T., Tang, X., Wang, J., & Zhao, L. (2025). Investigating shear force and torque of grapevine shoots based on experimental and simulation analysis. *BioResources*, *20*(3), 6662-6679. https://doi.org/10.15376/biores.20.3.6662-6679

Casado-García, 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. *Precision Agriculture*, *23*(6), 2001-2026. https://doi.org/10.1007/s11119-022-09929-9

Dong, Y., Hu, G., Liu, G., & Tohti, G. (2016). Segmentation method for grapevine critical structure based on Mask R-CNN model. *Journal of Chinese Agricultural Mechanization*, *45*(2), 207-214. https://doi.org/10.13733/j.jcam.issn.2095-5553.2024.02.030

Fernandes, M., Gamba, J. D., Pelusi, F., Bratta, A., Caldwell, D., Poni, S., … & Semini, C. (2025). Grapevine winter pruning: Merging 2D segmentation and 3D point clouds for pruning point generation. *Computers and Electronics in Agriculture*, *237*, 110589. https://doi.org/10.1016/j.compag.2025.110589

Gao, M., & Lu, T. F. (2006). Image processing and analysis for autonomous grapevine pruning. *International Conference on Mechatronics and Automation*, *2006*, 922–927. https://doi.org/10.1109/ICMA.2006.257748

Gentilhomme, T., Villamizar, M., Corre, J., & Odobez, J. (2023). Towards smart pruning: ViNet, a deep-learning approach for grapevine structure estimation. *Computers and Electronics in Agriculture*, *207*, 107736. https://doi.org/10.1016/j.compag.2023.107736

Guadagna, P., Fernandes, M., Chen, F., Santamaria, A., Teng, T., Frioni, T., Caldwell, D., Poni, S., Semini, C., & Gatti, M. (2023). Using deep learning for pruning region detection and plant organ segmentation in dormant spur-pruned grapevines. *Precision Agriculture*, *24*(4), 1547-1569. https://doi.org/10.1007/s11119-023-10006-y

Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., & Xu, C. (2020). *GhostNet: More features from cheap operations*. Paper presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR42600.2020.00165

Hou, Q., Zhou, D., & Feng, J. (2021). *Coordinate attention for efficient mobile network design*. Paper presented at the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR46437.2021.01350

Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., & Le, Q. (2019). *Searching for MobileNetV3*. Paper presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV.2019.00140

Hu, J., Shen, L., & Sun, G. (2018). *Squeeze-and-Excitation Networks*. Paper presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2018.00745

Íñiguez, R., Wolela, F., Gonzalez Pavez, M. I., Barrio, I., Tardáguila, J., Venter, T., & Poblete-Echeverria, C. (2025). Artificial intelligence-driven classification method of grapevine major phenological stages using conventional RGB imaging. *OENO One*, *59*(2). https://doi.org/10.20870/oeno-one.2025.59.2.9306

Karkee, M., Majeed, Y., & Zhang, Q. (2023). Advanced technologies for crop-load management. In S. G. Vougioukas & Q. Zhang (Eds.), *Advanced Automation for Tree Fruit Orchards and Vineyards* (p. 119-149). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-26941-7_6

Ma, N., Zhang, X., Zheng, H., & Sun, J. (2018). *ShuffleNetV2: Practical guidelines for efficient CNN architecture design*. Paper presented at the Computer Vision – ECCV 2018, Cham. https://doi.org/10.1007/978-3-030-01264-9_8

Majeed, Y., Karkee, M., & Zhang, Q. (2020b). Estimating the trajectories of vine cordons in full foliage canopies for automated green shoot thinning in vineyards. *Computers and Electronics in Agriculture*, *176*, 105671. https://doi.org/10.1016/j.compag.2020.105671

Majeed, Y., Karkee, M., Zhang, Q., Fu, L., & Whiting, M. D. (2019). A study on the detection of visible parts of cordons using deep learning networks for automated green shoot thinning in vineyards. *IFAC-PapersOnLine*, *52*(30), 82-86. https://doi.org/10.1016/j.ifacol.2019.12.501

Majeed, Y., Karkee, M., Zhang, Q., Fu, L., & Whiting, M. D. (2020a). Determining grapevine cordon shape for automated green shoot thinning using semantic segmentation-based deep learning networks. *Computers and Electronics in Agriculture*, *171*, 105308. https://doi.org/10.1016/j.compag.2020.105308

Marset, W. V., Pérez, D. S., Díaz, C. A., & Bromberg, F. (2021). Towards practical 2D grapevine bud detection with fully convolutional networks. *Computers and Electronics in Agriculture*, *182*, 105947. https://doi.org/10.1016/j.compag.2020.105947

McFarlane, N. J. B., Tisseyre, B., Sinfort, C., Tillett, R. D., & Sevila, F. (1997). Image analysis for pruning of long wood grape vines. *Journal of Agricultural Engineering Research*, *66*(2), 111-119. https://doi.org/10.1006/jaer.1996.0125

Mercurio, J. F., Gunkel, W. W., Sobel, T. A., Throop, J. A., & Norman, D. W. (1989). *Vision-guided block-type robotic grapevine pruner*. ASAE paper no. 89–7519, New Orleans, USA, 12–15 Dec.

Moreno, H., & Andújar, D. (2023). Proximal sensing for geometric characterization of vines: A review of the latest advances. *Computers and Electronics in Agriculture*, *210*, 107901. https://doi.org/10.1016/j.compag.2023.107901

Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z. (2023). *Efficient multi-scale attention module with cross-spatial learning*. Paper presented at the ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/ICASSP49357.2023.10096516

Rana, S., Gerbino, S., Akbari Sekehravani, E., Russo, M. B., & Carillo, P. (2024). Crop growth analysis using automatic annotations and transfer learning in multi-date aerial images and ortho-mosaics. *Agronomy*, *14*(9), 2052. https://doi.org/10.3390/agronomy14092052

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). *MobileNetV2: Inverted residuals and linear bottlenecks*. Paper presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2018.00474

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., & Hu, Q. (2020). *ECA-Net: Efficient channel attention for deep convolutional neural networks*. Paper presented at the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR42600.2020.01155