Segmentation and weight prediction of grape ear based on SFNet-ResNet18

Chang-Mei Lianga, Yan-Wen Liia, Yan-Hong Liia, Peng-Fei Wenb and Hua Yanga

aCollege of Information Science and Engineering, Shanxi Agricultural University, Taigu, Shanxi, People’s Republic of China; bCollege of Horticulture, Shanxi Agricultural University, Taigu, Shanxi, People’s Republic of China

ABSTRACT
In this paper, the segment and weight prediction problems are investigated for ear of grape based on deep learning technologies. The image datum is collected from ZaoHeiBao grape in a greenhouse by camera. The grape ear target segmentation model is constructed by cross combining three backbone networks (ResNet18, ResNet50, and ResNet101) and four deep learning semantic segmentation networks (SFNet, GCNet, EMANet, and Deeplabv3). The experimental results show that for the SFNet-ResNet18 model, whose structural size is 52.68MB, the mean Intersection over Union (mIoU) is 79.45%, the mean Pixel Accuracy (mPA) is 92.58%, and the average segmentation speed of the image (4000 × 3000) is 0.217s. Therefore, the performance of the SFNet-ResNet18 model outperforms other combined network models and is selected to segment grape ears. Furthermore, on the basis of the segmentation results of grape ears by using the SFNet-ResNet18 model, the grape ear weight is predicted by adopting fractional regression model. The value of $R^2$ is 0.8903, which means that the selected model could accurately predict the weight of grape ears. The proposed method can not only segment the grape ears and accurately predict the weight of the grape ears, but also provide theoretical and technical support for grape yield prediction.

1. Introduction
Grape (Vitis vinifera L.) is one of the four most widely cultivated fruits in the world which is rich in sugar, organic acid, amino acids, polyphenols and has physiological activities such as anti-cancer (J. He & Giusti, 2010; Xia et al., 2019). Since grape yield and fruit quality directly determine the interests of grape growers, rapid and accurate yield prediction has become one of the key problems in efficient cultivation (Gutiérrez-Gamboa et al., 2021). However, traditional grape yield forecasting is laborious and manual which makes it impossible to forecast quickly and accurately.

Grape target segmentation is the basis for predicting yield and constructing automated harvesting systems. In the early study of agricultural intelligence based on machine vision, morphological features such as colour, shape and texture of grape ears have been extracted for segmentation. Cao et al. have used grapes ear for recognition and segmentation under different spectral mapping (L. Cao et al., 2020). Based on the combination of LBP texture and colour to extract features, the KNN algorithm has been designed to segment red and white grape varieties by Badeka et al. (2019). Artificial neural network and genetic algorithm based on colour feature have been applied (Behroozi-Khazaei & MalekiM, 2017). Color mapping and morphological expansion techniques have been introduced to separate white and red grape varieties in a single cluster. Unfortunately, the segmentation using traditional machine vision technology relies too much on the quality of image acquisition. The real planting scene, lighting conditions, vine occlusion, grape ear attachment and other complex conditions bring great difficulties to high-quality image acquisition. In addition, the grape segmentation method based on feature engineering and machine learning requires tedious manual design, and its colour and geometric texture have a strong professional background. Therefore, the recognition accuracy and yield prediction accuracy in the real planting scene are limited.

In recent years, with the development of deep learning in various fields (Ding et al., 2018; Zhang et al., 2022), methods based on deep learning have been proved to be very effective in fruit target segmentation and yield prediction. The 10-layer convolutional neural network has been proposed by Muresan and Oltean (2017), Barth et al. (2019) to classify individual fruits such as harvested tomatoes, and has achieved good classification performance. Pixel-level segmentation of sweet
pepper has been trained by using a VGG network (Barth et al., 2018). By utilizing Faster R-CNN to fuse RGB and NIR image information, Deepfruits (Sa et al., 2016) has become a method for detecting fruits such as apples, mangoes, and sweet peppers. Furthermore, segmentation studies of cabbage (Jiang et al., 2018) and leaf (Ward et al., 2018) have been investigated by using Mask R-CNN. Grape particle detection based on a deep neural network has been proposed (Krabańek, 2018). A neural network classification method has been proposed for identifying grapes and estimating the yield (Aquino et al., 2017) which requires a single grape ear with a black background. Grimm et al. (2018) have taken VGG as the backbone network and have proposed an adaptive semantic segmentation network model to evaluate different growth stages and data sets of grape ears. Nelithimaru and Kantor (2019) have segmented grape ears using Mask R-CNN with 3D reconstructed images. Santos et al. (2020) have successfully detected instance level grape ears using Mask R-CNN. However, these methods are unable to detect blocked grape ears. Therefore, building an accurate and fast grape ear recognition and segmentation model is the key link of grape yield prediction, especially in the real growth scene.

In conclusion, under natural field conditions, the captured grape images are affected by various environmental conditions such as uneven illumination and complex backgrounds, which have a great impact on the real-time segmentation of grape images. Furthermore, it causes a large deviation in the prediction of grape yield. This study compares the target segmentation ability of different semantic segmentation models (Deeplabv3 (Chen et al., 2017), GCNet (Y. Cao et al., 2019), EMANet (Li et al., 2019), and SFNet (Li et al., 2020)). Meanwhile, in the enhanced feature extraction stage, different backbone networks are used to compare the ability of different feature layers to extract grape ear features, so as to obtain the optimal image features. By verifying and comparing the obtained 12 combined models on the grape ear test dataset, it is found that the SFNet-ResNet18 model has higher segmentation accuracy and speed. In addition, this study obtained a grape ear weight prediction model based on the effective segmentation of grape ear, and the $R^2$ value of the regression model is 0.8903.

The main contributions of this paper can be highlighted as follows: (1) The semantic segmentation model SFNet is improved by adding backbone feature extraction network ResNet18 in order to improve the feature extraction ability of the network. Meanwhile, the improved SFNet is validated on the newly collected grape bunches dataset and also compared against the Deeplabv3 and GCNet and EMANet with promising performance. (2) A power relationship is given between the semantic segmentation results and the weight of the grape spike.

2. Data acquisition and preprocessing

In this study, in order to achieve the grape spike target segmentation and yield prediction in real planting environment, according to the standard of experimental data collection, the grape images will be collected in real planting scene, compressed to $512 \times 512$, and inputted into the deep learning target segmentation model. The image segmentation results should be used to calculate the number of pixel, and the yield prediction value of grape will be obtained by the regression model.

2.1. Data acquisition

To achieve the grape spike target segmentation and weight prediction in real planting environment, both image and weight data of grape spike are needed to be collected. The image data was obtained from the grape greenhouse in Horticultural Experiment Station, Shanxi Agricultural University at 08:00–12:00 am on 14 September 2021. The weather was fine and the temperature was 15–20°C. The grape variety is ZaoHeiBao. The camera is Canon eosm6 with automatic exposure. Its exposure time is $1/200$ s and the shooting distance is about 300 mm. The shooting angle is parallel to the ground, and the collected image is about 12 million pixels $(4000 \times 3000)$. At the same time, the number of each grape spike image is numbered (for example, 1-2-8 indicates the grape spike of the first row, the second area and the serial number 8). The one-to-one correspondence between number and grape spike weight is recorded by weighing the grape spike.

After obtaining the segmented image data, it is necessary to label the sample data. In this paper, the interactive image segmentation method based on graph matching (Noma et al., 2012) is adopted, and the polygon shape is used to label the grape object accurately at the instance level. The results of different image annotation are shown in Figure 1. The data set does not carry out pruning, defoliation and any other intervention treatment on vineyard grape plants, which can reflect the real scene.

2.2. Data preprocessing

It is seen from Figure 1 that, in order to obtain the effective input of the model, the original data set is preprocessed in two steps to get the grape ear segmentation data set suitable for the real vineyard planting environment of this model.

(1) For the purpose of reducing the occupation rate of model display memory, reducing the amount of
computation, and accelerating the speed of model training, the original data set and its annotation results are continuously expanded and transformed, and the resolution of the transformed image is changed to $512 \times 512$.

(2) Considering the small number of training samples in the initial data set, data enhancement is performed on Figure 2(a) to enrich the data set and improve the generalization ability of the model. Each picture only performs 1–3 transformations randomly with 50% probability, including adding Gaussian noise, flipping $180^\circ$ and changing the brightness value to generate 4 enhanced pictures. The threshold modification range of brightness value is 0.5–1.5. Brightness value greater than 1 indicates dimming and less than 1 indicates brightening. The process is shown in Figure 2.

3. Model of grape fruit spike segmentation

3.1. The ResNet backbone network

The ResNet (K. He et al., 2016) network is often applied as a backbone network to extract the input image features. For the sake of solving the degradation problem in the deep network training process, the residual idea is introduced into the network structure, and the jump connection between deep network and shallow network is introduced, so as to avoid the saturation, even uneven decline of accuracy, and increase the number of network layers. In addition, ResNet series networks organize different feature graph sizes through stage during training, which can be used as effective input to subsequent networks. The results of feature extraction of the input images with different depths are greatly different. The depth of the ResNet network adopted in this paper is 18, 50 and 101 respectively (the depth of various networks refers to the number of network layers, such as convolutional layer, full connection layer, etc.). The specific structure of the three ResNets is shown in Table 1.

3.2. Semantic segmentation model

Semantic segmentation model generally consists of two parts: backbone network and task network. Full convolutional networks in backbone networks are subject to smaller effective perceptual domains and fail to adequately capture long-distance information. A series of deep learning modules proposed for the task network, especially the multi-scale context fusion module reflected in the task network, can effectively compensate for this defect. For example, Deeplab series model with Atrous Convolution pooling pyramid, GCNet network model with Global Attention map, EMANet network model with Expectation Maximum Attention Mechanism (EMA) and SFNet fused to Flow Alignment Module (FAM) and Pyramid Pooling Module (PPM).

Deeplabv3 designs serial and parallel pore-borne convolutional modules, using multiple different atrous rates to obtain multi-scale content information. Meanwhile, Deeplabv3 proposes the Atrous Spatial Pyramid Pooling (ASPP) module to mine convolutional features of different
scales and image layer features encoding global content information.

In order to understand the image from a global perspective as much as possible, GCNet integrates the advantages of NL Block’s strong global context modelling ability and SENet’s low computational cost. The Figure 3 shows the global context modelling framework (a), simplified version of NL Block (b), SE Block (c), and GC Block (d) proposed in the GCNet task network.

EMANet adopts the expectation maximization attention mechanism (EAM), which decreases the complexity by abandoning the calculation of the attention map on the whole image and iterating a set of compact bases through the expectation maximization algorithm as well as running the attention mechanism on this set of bases. EAM performs alternately and the feature plots are reconstructed after the convergence. EAM is embedded in the network to construct the semantic segmentation task network EMANet as shown in the Figure 4.

It is worth noting that the semantic flow module proposed by SFNet enables the deep semantic information to be transmitted to the shallow layer more effectively and facilitates the fusion of features of different layers. The SFNet structure resembles the FPN structure, where the PPM is the Pyramid Pooling Module and the FAM is the Flow Alignment Module. Each FAM structure will have two input features, in which the role is to assist in the fusion between these two features. Its network structure is shown in Figure 5.

### 3.3. Evaluation indicators

In this paper, the evaluation indicators commonly used in the target segmentation field includes Intersection over Union (IoU), mean Intersection over Union (mIoU), Pixel Accuracy (PA), and mean Pixel Accuracy (mPA).

IoU refers to the intersection of the predicted area and the actual area divided by the union of the predicted area and the actual area. The higher the IoU value is, the more the predicted region fits to the actual area, as shown in formula (1).

$$\text{IoU} = \frac{S_A \cap S_B}{S_A \cup S_B}$$  \hspace{1cm} (1)

where $S_A$ and $S_B$ represent the prediction box area pixel set and the real box area pixel set. The mIoU is the average of the IoU obtained from all test images.
PA refers to the proportion of pixels with correct category prediction in the total pixels. The higher the accuracy is, the better the model quality is shown in formula (2).

$$PA = \frac{\sum_{i=1}^{n} P_{ii}}{\sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}}.$$  \hspace{1cm} (2)

The mPA is average of the PA obtained from all image predictions.

3.4. Setting of the model training parameters

The test platform is configured as 11GB RTX2080Ti GPU and the system is Windows10 for model code programming based on the Paddleseg framework. Four pictures are trained in one batch. One round of iteration is to traverse all the training set data. In the experiment, the number of iteration rounds is set to 20. Image size is $512 \times 512$, the size of dataset before enhancement is 433, and the total size of dataset after data enhancement is 2165. The dataset is divided into training set, validation set and test set, where the size of training set is 1299 and the sizes of test set and validation set are 433. Using SGD as the optimizer and the learning rate of the optimizer is set to 0.02. To accelerate model convergence and avoid falling into local optimal solution or skipping optimal solution, the weight decay parameter is set to 0.0001 and the momentum is set to 0.9.

4. Results and analysis

4.1. Grape fruit ear segmentation and identification

As is known to all, backbone network, as a shared structure of various CNN models, is often suitable for feature pre-extraction operations. To explore the effect of different backbone networks on grape ear segmentation, three common RestNet-derived versions (ResNet18, ResNet50, ResNet101) are selected to extract features from the input images, and the extracted results are used as input to SFNet, EMANet, GCNet and Deeplabv3 task networks, respectively. Finally, the mIoU and mPA index values of
the segmentation on the test set are calculated. Table 2 shows the mIoU and mPA index values of grape ear segmentation by four task networks under the feature extraction results of three backbone networks on the test set and the model size and segmentation time output by different task networks under different backbone networks.

4.1.1. Performance effects of different backbone networks on grape spike segmentation

Different backbone networks differ effectively on grape spike segmentation in the SFNet, EMANet, GCNet, and Deeplabv3 task networks. For SFNet task network, ResNet18 as the backbone network is better than ResNet50 and ResNet101, with some improvements in mIoU and mPA indices. For GCNet and EMANet models, ResNet50 backbone network is the best. For Deeplabv3, ResNet101 backbone network is the best. As shown in Figure 6, in order to visually observe the effect of different backbone networks on task network model segmentation performance, this study plots a polyline graph with the loss function values of four task network models under different backbone networks. The smaller the segmentation loss value, the better the effect. The results demonstrate that under the condition of Deeplabv3 as the task network model, when ResNet101 is selected as the backbone network, the effect is better than that of ResNet18 and ResNet50. For the GCNet and EMANet models, the ResNet50 backbone network is optimal. For the SFNet task network, utilizing the ResNet18 as the backbone network is better than the ResNet50 and ResNet101. Various models can obtain low loss values from the eighth round of iteration, which means that different backbone networks can update the model weight well in the early stage of model training to make the model tend to the optimal value. At the same time, it is not difficult to find that for the same task network, the selection of backbone network affects the segmentation effect. Compared with ResNet18, ResNet50 and ResNet101 networks are deeper and can theoretically extract more complex local information. However, for specific task-related networks, a deeper backbone network is more beneficial to the task. In fact, the appropriate depth of the backbone network should be selected according to the data and task characteristics.

4.1.2. Performance effects of different task networks on grape spike segmentation

The same backbone network adopts different task networks, and the grape segmentation effect is different. Figure 7 displays the change curve of the training loss function of each backbone network under different task networks. Combined with the mIoU and mPA indicators in Table 2, it shows that when employing different backbone networks, the segmentation accuracy of SFNet as
Table 2. Status of grape spike segmentation index values for different task networks under different backbone networks.

| Task networks | BackBone   | mIoU   | mPA   | Times (ms/step) | Model Size (KB) |
|---------------|------------|--------|-------|-----------------|-----------------|
| SFNet         | ResNet18   | 0.7945 | 0.9258| 211             | 53944           |
|               | ResNet50   | 0.7918 | 0.9522| 277             | 164215          |
|               | ResNet101  | 0.7900 | 0.9242| 235             | 238634          |
| EMA Net       | ResNet18   | 0.7927 | 0.9247| 222             | 131305          |
|               | ResNet50   | 0.7942 | 0.9253| 227             | 165784          |
|               | ResNet101  | 0.7926 | 0.9238| 236             | 240203          |
| GCNet         | ResNet18   | 0.7820 | 0.9209| 219             | 81797           |
|               | ResNet50   | 0.7833 | 0.9213| 230             | 194078          |
|               | ResNet101  | 0.7739 | 0.9176| 241             | 268496          |
| Deeplabv3     | ResNet18   | 0.7577 | 0.9130| 213             | 59963           |
|               | ResNet50   | 0.7743 | 0.9182| 225             | 152847          |
|               | ResNet101  | 0.7745 | 0.9192| 232             | 227265          |

Figure 6. Training loss function change curve for each task network model under different backbone networks.

a segmentation task network is substantially better than the other three networks. Hence, the selection of the task network is of great significance for the further improvement of the segmentation performance. In the end, the SFNet-ResNet18 model is adopted in this paper to obtain a better segmentation effect, and the average segmentation accuracy reaches 92.58%.

In addition, comparing the model size and segmentation speed of different combined models, it can be seen that under the three backbone networks, compared with the other 11 models, the SFNet-ResNet18 model has a minimum size of 52.67MB, and the segmentation speed is the fastest, with an average of 0.211 s per image. Generally, the larger the model, the slower the segmentation speed. But it does not mean that the larger the model, the higher the segmentation accuracy. Only Deeplabv3 presents the feature that the larger the model, the higher the accuracy. For GCNet and EMA Net, Using the ResNet50 backbone network achieves the highest segmentation accuracy, while for the SFNet task network, the ResNet18 backbone network has the highest segmentation accuracy.

4.1.3. Visualization of grape fruit spike segmentation results

On the other hand, we divided the test set into four scenarios: Frontlighting, Backlighting, Paste and Shelter in order to further study the robustness of the model in different environment. With ResNet18 as the backbone network, the partial visualization results are shown in Figure 8.
Figure 7. Change curve of the training loss function for each backbone network under different task networks.

4.2. Grape production forecast

The standard grape yield estimation is given by the following formula (3).

$$Y = N_v \times N_b \times P_b.$$  (3)

where $N_v$, $N_b$ and $P_b$ represent the number of vines per surface unit, the number of grape spikes per vine and the average weight of grape spikes, respectively.

Grape yield is obtained by examining the number of grape ears and the average spike weight in the vine. Among them, the detection of grape ears has a certain scientific and theoretical support in computer vision tasks. The average weight of grape ears can only be accurately determined near the harvest period, and it significantly changes from year to year, which is difficult and inaccurate to estimate based on historical data alone. Therefore, this study is committed to using deep learning target segmentation method for accurate and efficient real-time segmentation, providing a basis for further prediction of grape ear weight.

4.2.1. Statistics of grape ears

The premise of accurate statistic of grape spike number is Grape spike detection which can be regarded as the work of semantic segmentation task. In this study, we will classify and sort out with the method of grape string detection. Among them, Santos et al. (2020) provides the training of grape spike detection dataset for deep learning models, and uses advanced detection models such as YOLOv2 (Redmon et al., 2016), YOLOv3 (Redmon & Farhadi, 2018), and Mask R-CNN (K. He et al., 2017) for the WGISD dataset. Sozzi et al. (2021) also applies YOLOv4 (Bochkovskiy et al., 2020) to verify the effectiveness of deep learning object detection method migration in the field of grape string detection.

4.2.2. Weight prediction of grape fruit spike

Grape spike weight prediction is a further work of grape spike target segmentation. In order to verify the effectiveness of the deep learning segmentation network SFNet-ResNet18 model for predicting the weight of grape ears, this study numbers and picks the grape ears in the above segmented image data, and weighs and records the weight in sequence according to the number. The weight value is accurate to 1.0 g, this experiment finally obtains 433 grape ear weights as test samples. The grape spike weight prediction model is constructed with the number of pixels in the segmented image as the independent variable and the actual weight of the grape spike as the dependent variable to obtain the regression model.

First, It can be seen that the Pearson correlation coefficient between the number of pixels in the segmented image and the actual weight of the grape ear is 0.9381 through the correlation analysis between the actual weight of grape spike and the pixel number of the segmented image. Therefore, the segmented pixel number can be determined as the predicted weight of grape spike.

Since the grape berry approximates the spheres, there is a power relationship between the cross section area and the grape spike weight expressed by the number of pixels, and a linear relationship between $3/2$ power of the cross section area and weight, assuming the established grape spike weight prediction model is formula (4).

$$W = aS^b + c.$$  (4)

where $W$ is the weight of grape ear with the unit g, $S$ is the cross-sectional area with the unit cm$^2$. $a$, $b$ and $c$ are the estimated parameters.

After fitting the data with R software, the fit curves are shown in Figure 9 and the regression model is shown in...
Figure 8. ResNet18 shows the visualization of grape spike segmentation for different network models under the backbone network. (a) Frontlighting; (b) Backlighting; (c) Paste; (d) Shelter.

The $R^2$ value of the regression model is 0.8903 and the $P$ value is less than 0.0001, indicating that this model could more accurately detect grape ear weight.

After the regression model is obtained, the grape ear weight in the real planting scene can be predicted: firstly, the grape image data in the real planting scene can be obtained according to the standard of experimental data collection, and the image whose size is compressed to $512 \times 512$ is input into the deep learning target segmentation model. Then, the image segmentation result of grape ear is obtained. Next, the pixel points of grape ear segmentation are calculated. Finally the regression model is introduced to obtain the weight prediction value of grape ear.

5. Conclusion
The paper focuses on analysing the application of network segmentation model in different tasks of grape yield...
prediction from two aspects of segmentation accuracy and time validity, and draws the following conclusions:

(1) Compared with other combined models, the SFNet-ResNet18 combined model is the smallest, with a model size of 52.67 MB. It has the fastest segmentation speed, with an average segmentation speed of 0.211 s per image. Its segmentation accuracy is the highest, and the average accuracy mPA value is 92.58%, indicating that the SFNet-ResNet18 combined model can efficiently and accurately segment the fruit ears of ZaoHeiBao grapes.

(2) ZaoHeiBao grapes berry approximates the sphere, and the SFNet-ResNet18 combination model has a 3/2 power relationship between the number of pixels in the grape spike target segmentation and the actual grape spike weight. The $R^2$ value obtained by introducing nonlinear fitting is 0.8903, and the error between the cross-sectional area power 1.4872 and theoretical power 1.5 is only 0.0128, which proves that the nonlinear regression model obtained in this paper can accurately predict grape ear weight.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Funding**

This article is funded by the Basic Research Program of Shanxi Province [grant numbers 20210302123408, 201901D111222 and 202103021223141].

**References**

Aquino, A., Diago, M. P., Millán, B., & Tardáguila, J. (2017). A new methodology for estimating the grapevine-berry number per cluster using image analysis. *Biosystems Engineering, 156*, 80–95. https://doi.org/10.1016/j.biosystemseng.2016.12.011

Badeka, E., Kalabokas, T., Tziridis, K., Nicolaou, A., Vrochidou, E., Mavridou, E., & Pachidis, T. (2019). Grapes visual segmentation for harvesting robots using local texture descriptors. In *International conference on computer vision systems* (pp. 98–109). Springer.

Barth, R., IJsselmuiden, J., Hemming, J., & Van Henten, E. J. (2018). Data synthesis methods for semantic segmentation in agriculture: A capsicum annum dataset. *Computers and Electronics in Agriculture, 144*, 284–296. https://doi.org/10.1016/j.compag.2017.12.001

Barth, R., IJsselmuiden, J., Hemming, J., & Van Henten, E. J. (2019). Synthetic bootstrapping of convolutional neural networks for semantic plant part segmentation. *Computers and Electronics in Agriculture, 161*, 291–304. https://doi.org/10.1016/j.compag.2017.11.040

Behroozi-Khazaei, N., & MalekiM, M. R. (2017). A robust algorithm based on color features for grape cluster segmentation.
Krabánek, P. (2018). Deep grapes: Precise detection of grapes in low-resolution images. *IFAC-Papers On Line*, 51(6), 185–189. https://doi.org/10.1016/j.ifacol.2018.07.151

Li, X., You, A., Zhu, Z., Zhao, H., Yang, M., Yang, K., Tan, S., & Tong, Y. (2020). Semantic flow for fast and accurate scene parsing. In ECCV (pp. 775–793).

Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., & Liu, H. (2019). Expectation-maximization attention networks for semantic segmentation. In ICCV (pp. 9167–9176).

Muresan, H., & Oltean, M. (2017). Fruit recognition from images using deep learning. ar Xiv preprint arXiv:1712.00580. https://doi.org/10.48550/arXiv.1712.00580.

Nellithimaru, A. K., & Kantor, G. A. (2019). ROLS: Robust object-level SLAM for grape counting. In *Proceedings of the international conference on computer vision and pattern recognition workshops*.

Noma, A., Graciano, A. B., Cesar Jr, R. M., Consularto, L. A., & Bloch, I. (2012). Interactive image segmentation by matching attributed relational graphs. *Pattern Recognition*, 45(3), 1159–1179. https://doi.org/10.1016/j.patcog.2011.08.017

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the international conference on computer vision and pattern recognition* (pp. 779–788). https://doi.org/10.1109/CVPR.2016.91.

Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. Co RR.. https://doi.org/10.48550/arXiv.1804.02767.

Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T., & McCool, C. (2016). Deepfruits: A fruit detection system using deep neural networks. *Sensors*, 16(8), 1222. https://doi.org/10.3390/s16081222

Santos, T. T., De Souza, L. L., Dos Santos, A. A., & Avila, S. (2020). Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Computers and Electronics in Agriculture*, 170, 105247. https://doi.org/10.1016/j.compag.2020.105247

Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., & Marinello, F. (2021). Grape yield spatial variability assessment using YOLOv4 object detection algorithm. In *Precision agriculture* (pp. 193–198). Wageningen Academic Publishers. https://doi.org/10.3920/978-90-8686-916-9-22

Ward, D., Moghadam, P., & Hudson, N. (2018). Deep leaf segmentation using synthetic data. ar Xiv preprint arXiv:1807.10931. https://doi.org/10.48550/arXiv.1807.10931.

Xia, S., Qiao, X., Chen, H., Nan, H., & Zhang, Z. (2019). Coordinated regulation of grape berry flesh color by transcriptional activators and repressors. *Journal of Agricultural and Food Chemistry*, 67(42), 11815–11824. https://doi.org/10.1021/acs.jafc.9b05234

Zhang, K., Wang, C., Yu, X., Zheng, A., Gao, M., Pan, Z., Chen, G., & Shen, Z. (2022). Research on mine vehicle tracking and detection technology based on YOLOv5. *Systems Science & Control Engineering*, 10(1), 347–366. https://doi.org/10.1080/21642583.2022.2057370