







# Dual Attention-Guided Network for Anchor-Free Apple Instance Segmentation in Complex Environments

Yunshen Pei<sup>1</sup>, Yi Ding<sup>2</sup>, Xuesen Zhu<sup>1</sup>, Liuyang Yan<sup>1</sup>,  
and Keyang Cheng<sup>1</sup>

<sup>1</sup> School of Computer Science and Communication Engineering, Jiangsu University,  
Zhenjiang 212013, China

[kycheng@ujs.edu.cn](mailto:kycheng@ujs.edu.cn)

<sup>2</sup> School of Electronic Science and Engineering, Nanjing University,  
Nanjing 210023, China

[181180019@smail.nju.edu.cn](mailto:181180019@smail.nju.edu.cn)

**Abstract.** Apple segmentation is an important part of the automatic picking system of apple plantation. However, due to the complexity of apple orchard environments, including light change, branch and leaf occlusion and fruit overlap, the segmentation accuracy of the existing methods is limited, which affects the large-scale application of the automatic picking system. To solve these problems, this paper proposes a new apple instance segmentation method based on a dual attention-guided network. Firstly, the image is preprocessed by the Image Correction Module (ICM) to improve the robustness of the network to the natural environment. Secondly, the Multi-Scale Enhanced Fusion Feature Pyramid Network (MSEF-FPN) is used as the feature extraction module to enhance the ability of image feature extraction, so as to reduce the interference of complex background on apple instance segmentation results without increasing the amount of calculation. Then, a new Dual Attention-Guided Mask (DAGM) branch is added to focus on the pixels of irregular occlusion and overlapping objects, and accurate pixel-level mask segmentation is carried out in the detection rectangular bounding box. Finally, this study carried out instance segmentation experiments on apples with different lighting conditions and different occlusion. The test results show that the model proposed in this paper has excellent detection accuracy, robustness and real-time, and has important reference value for solving the problem of accurate fruit recognition in complex environments.

**Keywords:** Apple segmentation · Complex environments · Feature extraction

## 1 Introduce

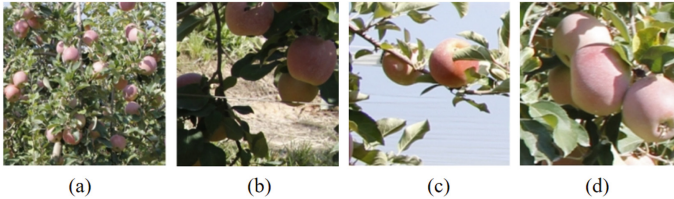
With the wide application of deep learning in the field of computer vision [1, 2], increasingly researchers are engaged in intelligent agriculture-related work.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

S. Yu et al. (Eds.): PRCV 2022, LNCS 13537, pp. 533–545, 2022.

[https://doi.org/10.1007/978-3-031-18916-6\\_43](https://doi.org/10.1007/978-3-031-18916-6_43)

At present, practical problems in the agricultural production processes (such as pest prediction and monitoring, automatic harvest, etc.) have been solved by many network model methods. However, with the improvement in operating efficiency and upgrading of agricultural equipment, the requirements for real-time performance and accuracy of operating machines have also gradually increased [3–5], and the requirements for agricultural equipment vision system are also higher and higher [6, 7]. In the complex orchard environment, the results of target fruit detection [8, 9] and segmentation [10–12] limit the performance of the visual system. Such as apple density (Fig. 1(a)), illumination angle change (Fig. 1(b)), branch and leaf occlusion (Fig. 1(c)) and overlapping apple (Fig. 1(d)) will have a certain influence on target detection, which brings great difficulties and challenges to the accurate recognition of fruits.



**Fig. 1.** Illustration of our framework. (a) Apple density; (b) illumination angle change; (c) branch and leaf occlusion; (d) overlapping apple

To address these deficiencies while considering the above factors, an effective and accurate Apple Instance Segmentation method based on a dual attention-guided network is proposed to improve the segmentation accuracy of apple in complex environments. More precisely, the main contributions of this paper are summarized as follows:

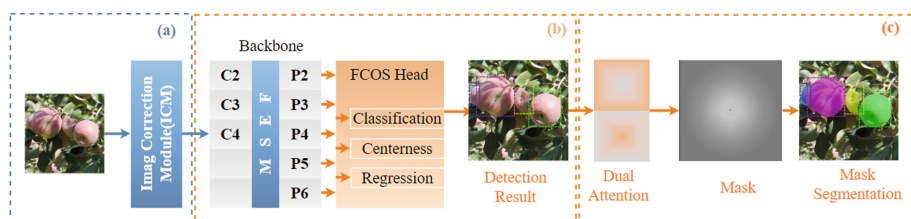
- (1) Aiming at the problems of illumination, occlusion and overlap in a complex environments, an anchor-free apple instance segmentation method based on dual attention-guided.
- (2) An adaptive Image Correction Module (ICM) is introduced to enhance the robustness of the image to natural illumination and contrast changes.
- (3) The Multi-Scale Enhancement Fusion (MSEF) module is introduced into the feature pyramid network (FPN). Its purpose is to enhance the feature extraction ability of the image and reduce the interference of complex background to apple detection results without increasing the amount of calculation.
- (4) To improve the segmentation accuracy of overlapping and occluding apples in complex environments, a new branch of Dual Attention-Guided Mask (DAGM) is added to deal with fruit occlusion and overlap.

## 2 Related Works

Traditional machine learning methods have made important contributions to fruit detection and segmentation [13–15]. A yield prediction strategy based on texture, fruit color and edge shape was proposed, and the recognition rate of green apples under natural light was close to 95% [16]. Tian used RGB spatial information to locate the center and radius of the apple, and combined with depth image information to match the target area [17]. These methods are not sufficient to identify overlapping or clustered fruits. To solve the above problems, a robust apple image segmentation algorithm based on a fuzzy reasoning system, which improves the generalization ability of segmentation [18]. However, due to the lack of in-depth analysis of image features, the above methods are usually poor in robustness and adaptability in complex environments with occlusion and overlapping.

With the rapid development of deep learning theory, an increasing number of deep learning methods have been proposed for agricultural fruit detection [19–24]. Liu used the improved single-stage detector Yolo-V3 to conduct tomato positioning detection in complex scenes [25]. Jia combined DenseNet and ResNet as the feature extraction backbone of the original model to improve the Mask R-CNN, greatly improving the identification accuracy of Apple in the overlapping and occlusion environments [4]. Compared with traditional visual methods, the accuracy and applicability of the recognition model based on deep learning have been greatly improved. However, these methods usually require many computing and storage resources, which will seriously affect the segmentation speed and operation stability of agricultural equipment in practical applications.

## 3 Methods



**Fig. 2.** Illustration of our framework. (a) Image Correction Module (ICM); (b) Anchor-free Detection Module; (c) Dual Attention-guided Mask (DAGM) Module

To improve the accuracy and efficiency of apple instance segmentation in complex environments, an accurate and efficient anchor-free apple instance segmentation method based on dual attention-guided network. The framework of our method is

shown in Fig. 2, which includes three parts: (a) Image Correction Module (ICM); (b) Anchor-free Detection Module; (c) Dual Attention-guided Mask (DAGM) Module.

### 3.1 Image Correction Module (ICM)

To cope with the challenge of illumination change, ICM is used to transform images under different illuminations into similar illumination. The module follows IBNet [26] and realizes image adaptive correction by constructing an encoding and decoding network (as shown in Fig. 3). First, the convolutional neural network is used to extract image features, and then deconvolution is used as a decoder for resampling to restore and correct the image with the same size as the input image. In the deconvolution process, network parameters are trained to ensure that the corrected images have similar illumination intensities.

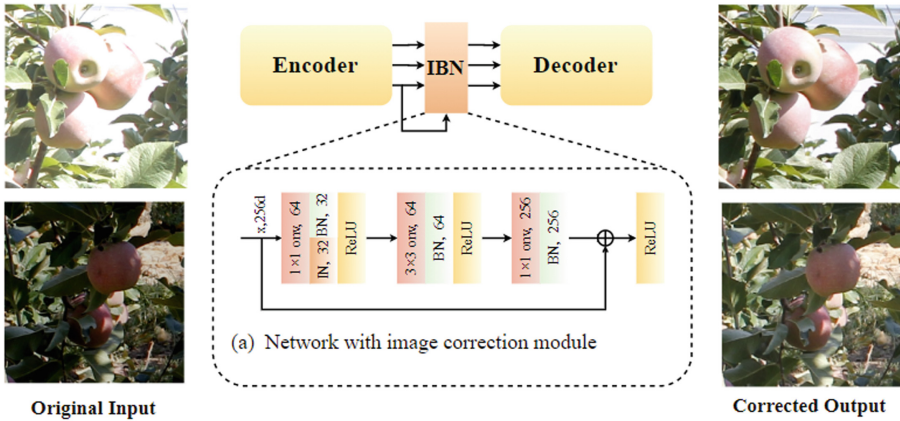
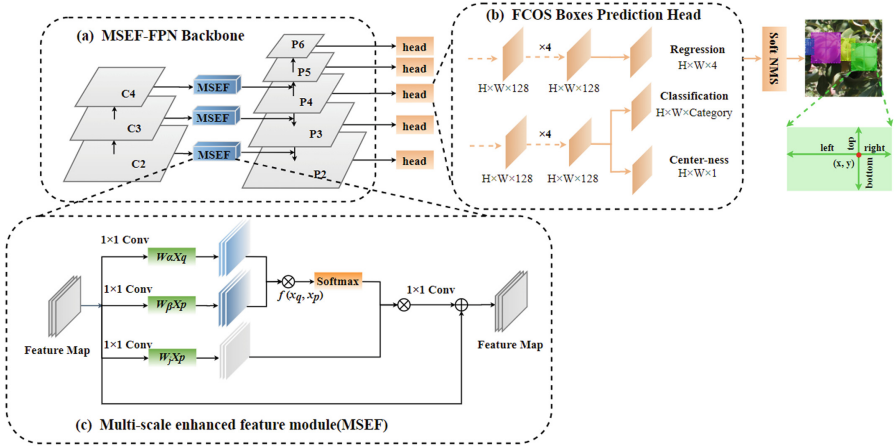


Fig. 3. Schematic diagram of the image correction module (ICM).

ICM is built on a trainable network, allowing for end-to-end training and adaptive correction of images rather than increasing or decreasing brightness at a specific rate during data enhancement. The basic network structure of ICM is shown in Fig. 3(a). The underlying features of convolutional neural network reflect the appearance features of objects, such as texture and color, while the high-level features reflect the semantic information of the target. Therefore, by adding image normalization in the lower layer of the network, the distribution of image data under different illumination can be adjusted to a similar area, increasing the adaptability of the network to illumination, and thus reducing the influence of illumination changes.



**Fig. 4.** Architecture of bounding box pre detection. (a) MSEF-FPN Backbone; (b) FCOS Boxes Prediction Head; (c) Multi-scale enhanced feature module (MSEF).

### 3.2 Feature Extraction MSFM-FPN Detection Network

Our model uses FCOS (shown in Fig. 4) as the basic detection model and improves it. The introduction of MSFM into the lateral connection in FPN solves some defects of FPN. For example, direct fusion of these features may be reduce the representation of multi-scale features due to the inconsistency of semantic information. In addition, in the process of picking, due to the interference of shooting distance, occlusion, or overlap, the proportion of the target in the image is insignificant. After the deep convolution operation on the image, the target feature map will be changed to a small extent, which greatly reduces the spatial information contained in the feature map, thus reducing the detection accuracy. Therefore, to gather multi-scale features and maintain a high-resolution representation in the process of convolution, MSFM is introduced at the lateral connection of FPN to improve the feature extraction capability of the image. Figure 4(c) shows the overall content of MSFM after improvement.

First, we use two weight transformations  $W_\alpha X_q$  And  $W_\beta X_p$  to reduce the number of channels and then reduce the amount of calculation. Multiply the two output matrices (where  $W_\alpha X_q$  will be transposed), calculate the similarity, and then perform the softmax operation to obtain the position attention, that is, the normalized correlation between each pixel in the current feature map and all other position pixels. Finally, by multiplying with  $W_j X_p$  matrix, the position attention mechanism is applied to the corresponding position of each feature graph of all channels. Restore the output channel through  $1 \times 1$  convolution to ensure that the input and output scales are exactly the same. The corresponding nonlocal operations are shown in Eqs. (1), (2) and (3).

$$f(x_q, x_p) = e^{(W_\alpha X_q)^T (W_\beta X_p)} \quad (1)$$

$$C(x) = \sum_{\forall p} f(x_q, x_p) \tag{2}$$

$$y_q = \frac{1}{C(x)} f(x_q, x_p)(W_j X_p) = softmax((W_\alpha X_q)^T (W_\beta X_p)(W_j X_p)) \tag{3}$$

### 3.3 Dual Attention-Guided Network for Instance Segmentation

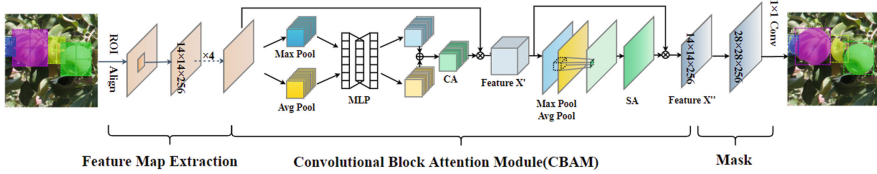


Fig. 5. The architecture of DAGM.

Considering that fruits are located in complex environments, many interference factors greatly reduce the segmentation effect of the model. Therefore, we designed a Dual Attention-guided Mask module (DAGM), as shown in Fig. 5. Compared with the general segmentation framework, this method has obvious advantages in the segmentation of small objects with serious occlusion or overlap. At the front-end of the framework, the convolutional network is used for feature extraction, and at the back end, conditional/Markov random fields are used to optimize the front end output, and the segmentation results are obtained.

The DAGM branch applies the boundary box predicted by FCOS to further predict the segmentation mask of each region of interest (ROI). Firstly, the predicted ROI is distributed to different FPN feature layers according to the resolution, and ROI Align is used for feature alignment. This is similar to using the Mask R-CNN to predict the segmentation mask. However, the relationship between the original image resolution and the ROI size must be considered in order to reasonably allocate the ROI to the feature layer of the corresponding resolution (considering the FPN multi-scale strategy). Secondly, after extracting the features in ROI with  $14 \times 14$  resolution in ROI Align, these features are transmitted to Convolutional Block Attention (CBA) network. Specifically, the characteristics are divided into maximum pool and average pool to obtain two groups  $1 \times 1 \times C$  characteristic matrix and transfer it to MLP, and then add the two output characteristic matrices to obtain the weight information Channel Attention (CA) of different channels. The calculation method is shown in Eq. (4). After CA is multiplied by the input characteristic matrix, the characteristic matrix combined with channel attention is obtained, as shown by Feature X' in Fig. 5. Then the feature matrix fused with channel attention is passed through  $W \times H \times 1$ , and condense the two feature maps in the depth direction, and then perform convolution operation to obtain spatial attention (SA) integrating spatial weight information. The calculation method is shown in Eq. (5).

In Eq. (5),  $f^{7 \times 7}$  indicates that the size of the pooling kernel is  $7 \times 7$ . Finally, SA is multiplied by feature X' to obtain the feature map Refined Feature X", which combines channel and spatial attention information.

$$CA(X) = \Sigma(MLP(maxPool(X)) + MLP(avgPool(X))) \quad (4)$$

$$SA(X) = \sigma(f^{7 \times 7}([maxPool(X'); avgPool(X')])) \quad (5)$$

Then, the obtained enhanced spatial attention feature map is up sampled to generate a feature map with a resolution of  $28 \times 28$ . The  $1 \times 1$  convolution kernel is used to generate the mask of instance segmentation.

### 3.4 The Loss Functions

The overall loss function  $L_{total}$  (as shown in Eq. (6)) of the model is composed of  $L_{cls}$ ,  $L_{reg}$ ,  $L_{center}$  and  $L_{mask}$ , where  $L_{cls}$  is the classification loss,  $L_{center}$  is the center-ness loss,  $L_{reg}$  is the box regression loss, and  $L_{mask}$  is the mask loss using the average binary crossentropy loss

$$\begin{aligned} L_{total} = & \frac{1}{N_{pos}} \sum_{x,y} L_{cls}(p_{x,y}, p_{x,y}^*) + \frac{\lambda}{N_{pos}} \sum_{x,y} p_{x,y}^* L_{reg}(d_{x,y}, d_{x,y}^*) \\ & + \frac{\beta}{N_{pos}} \sum_{x,y} p_{x,y}^* L_{center}(center_{x,y}, center_{x,y}^*) + L_{mask}(s_x, s_x^*) \end{aligned} \quad (6)$$

In Eq. (6),  $p_{x,y}$ ,  $d_{x,y}$  and  $center_{x,y}$  are the predicted values of classification branch, regression branch and centrality branch at the spatial position  $(x, y)$ .  $p_{x,y}^*$ ,  $d_{x,y}^*$  and  $center_{x,y}^*$  correspond to the training target at the spatial position  $(x, y)$ . Among the three loss items,  $L_{reg}$  and  $L_{center}$  are only for positive samples,  $N_{pos}$  is the number of positive samples, and  $\lambda$  and  $\beta$  are the balance coefficients of each loss item.

The classification loss  $L_{cls}$  in Eq. (6) is shown in Eqs. (7) and (8):

$$L_{cls}(p_{x,y}, p_{x,y}^*) = -\alpha_t (1 - p_{x,y}^t)^\gamma \log(p_{x,y}^t) \quad (7)$$

$$p_{x,y}^t = \begin{cases} p_{x,y} & \text{if } p_{x,y}^* = 1 \\ 1 - p_{x,y} & \text{otherwise,} \end{cases} \quad \alpha_t = \begin{cases} \alpha & \text{if } p_{x,y}^* = 1 \\ 1 - \alpha & \text{otherwise} \end{cases} \quad (8)$$

where  $\alpha$  Responsible for balancing the importance between positive and negative samples,  $\gamma$  responsible for adjusting the rate of weight reduction of simple samples.

The regression loss  $L_{reg}$  in Eq. (6) is shown in Eq. (9):

$$L_{reg}(d_{x,y}, d_{x,y}^*) = -\ln \frac{Intersection(d_{x,y}, d_{x,y}^*)}{Union(d_{x,y}, d_{x,y}^*)} \quad (9)$$

where  $intersection(d_{x,y}, d_{x,y}^*)$  and  $Union(d_{x,y}, d_{x,y}^*)$  are the intersection area and combined area between the prediction frame and the real frame respectively.

The center-ness loss  $L_{center}$  in Eq. (6) is shown in Eq. (10):

$$L_{center}(center_{x,y}, center_{x,y}^*) = -(center_{x,y} \log(center_{x,y}^*) + (1 - center_{x,y}) \log(1 - center_{x,y}^*)) \quad (10)$$

The mask loss  $L_{mask}$  in Eq. (6) is shown in Eq. (11):

$$L_{mask} = \sum_x -[s_x^* \log(s_x) + (1 - s_x^*) \log(1 - s_x)] \quad (11)$$

where  $s_x$  is the probability that the x-th pixel belongs to the target pixel and  $s_x^*$  is the probability that the x-th pixel belongs to the real target pixel.

## 4 Experiment

### 4.1 Dataset and Evaluation Metrics

**Apple Dataset Acquisition.** In this paper, we choose the open dataset Fuji SFM dataset, and make appropriate modifications to the data set to cooperate with the experiment of this paper. We select 400 appropriate Apple images from 582 images (the resolution of each image is  $5184 \times 3456$ ), then cut 15 images with the resolution of  $1024 \times 1024$  from each image, and get 6000 images with the resolution of  $1024 \times 1024$ . Then select the appropriate 1400 images from the 6000 images as the final data set. Finally, in order to make the network model have high accuracy and robustness, we use the mainstream image annotation tool labelme to annotate and store the data set manually. Figure 1 shows some images in the dataset.

**Evaluation Metrics.** We follow the internationally unified measurement standards and use the  $AP$  (average precision),  $AP_{50}$  ( $AP$  for IoU threshold 50%) and  $AP_{75}$  ( $AP$  for IoU threshold 75%) to measure the quality of the model.

### 4.2 Implementation Details

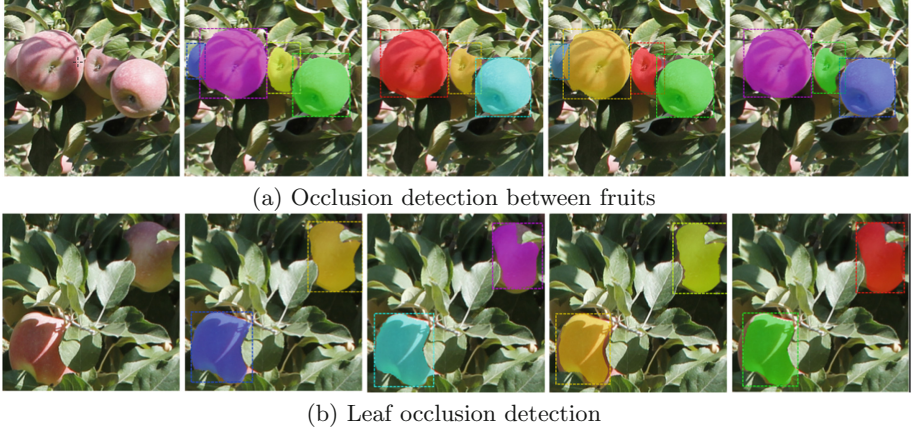
In the training stage, we trained a total of 50 epochs with 200 steps each. And the initial learning rate is 0.01. The Network parameters are also optimized using adaptive moment estimation (Adam). The momentum, as well as decay weights, are 0.9 and 0.0001, respectively.

### 4.3 Comparative Experiments

In this study, apple fruits with different occlusion and different lighting conditions were detected on the computer workstation. The detection effects of Mask R-CNN, SOLO [27], PolarMask [28] and our method under the above conditions were compared, and the performance of the algorithm was evaluated with  $AP$ ,  $AP_{50}$ ,  $AP_{75}$ .



**Comparison Experiment of Overlapping and Branch and Leaf Occlusion.** In the natural environment, there will be overlapping fruits and fruits covered by branches and branches. The contour information of the fruit part is lost, which increases the difficulty of fruit detection. Therefore, this study tested the overlapping of fruits and different degrees of branch and stem shielding. The statistical results are shown in Fig. 6 and Table 1.

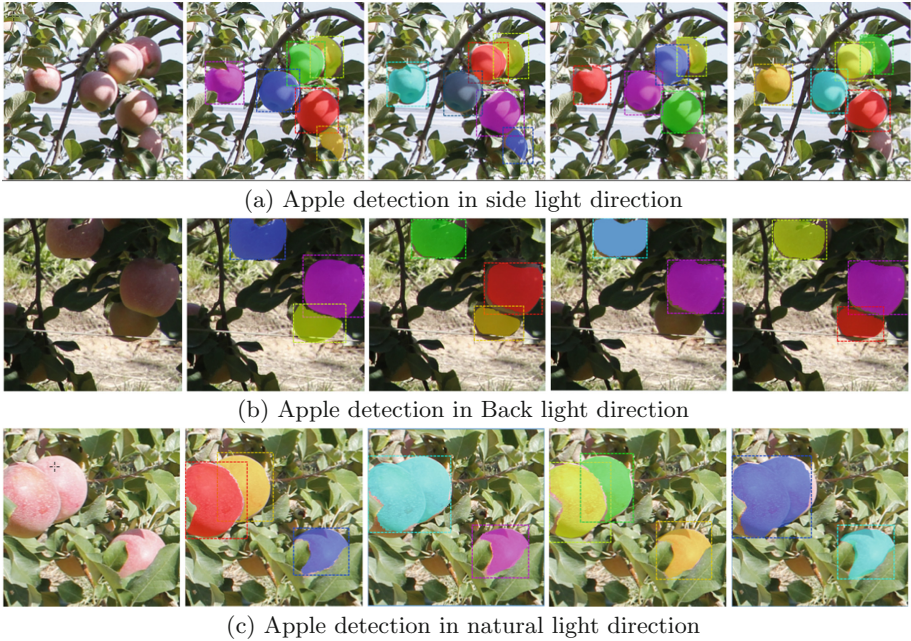


**Fig. 6.** Detection effect of 4 algorithms on different occluded apples. (From left to right, there are pictures of the detection results of Original Image, our Model, Mask R-CNN, SOLO and Polar Mask)

**Table 1.** Experimental results of 4 algorithms for images with different occluded apples.

Occlusion category	Algorithm	$AP$	$AP_{50}$	$AP_{75}$
Apple overlap	Mask R-CNN	83.2	72.1	64.7
	SOLO	86.3	75.7	67.8
	PolarMask	86.8	75.5	68.2
	Ours	<b>88.1</b>	<b>77.2</b>	<b>69.3</b>
Leaf occlusion	Mask R-CNN	81.2	70.3	62.7
	SOLO	83.6	74.9	66.8
	PolarMask	84.8	74.7	64.2
	Ours	<b>85.1</b>	<b>76.8</b>	<b>68.6</b>

As can be seen from Table 1, in the apple overlapping scenario, the  $AP$  value of our algorithm is 4.9%, 5.1% and 4.6% higher than that of Mask R-CNN respectively. The  $AP$  value of the latter two algorithms in both cases is lower than that of the algorithm in this paper. From the comprehensive results, our algorithm can be competent in the detection of different occlusion and overlap.



**Fig. 7.** Detection effect of 4 algorithms on apples under different illumination. (From left to right, there are pictures of the detection results of Original Image, our Model, Mask R-CNN, SOLO and Polar Mask)

**Comparison Experiment with Different Light.** Under the conditions of natural light, back light and side light, the fruit will be brighter or darker. And due to the great influence of dense apple samples, dense apple samples will not be considered when selecting images here. The statistical results are shown in Fig. 7 and Table 2.

As can be seen from Table 2, the  $AP$  value of the improved model in three different scenarios is higher than that of the other three algorithms. From different scenes, the four algorithms perform best in side light, while the model performs worst in backlight. Because the texture of the apple is clear under side light, the surface illumination intensity is uniform, and the backlight condition will cause some interference to the detection. Overall, our model can adapt to the influence of lighting conditions on apple surface color, texture features and contour, and can effectively detect apples in complex images.

#### 4.4 Ablation Experiment

In this section, in order to clarify the impact of image correction module (ICM) and multi-scale enhanced fusion feature pyramid module (MSEF-FPN) on the performance of the model, ablation research is carried out, and the role of each module is analyzed in detail. We gradually introduced our module, tested the  $AP$

**Table 2.** Experimental results of 4 algorithms for apples under different illumination.

Illumination angles	Algorithm	$AP$	$AP_{50}$	$AP_{75}$
Side light	Mask R-CNN	88.9	80.2	72.7
	SOLO	90.2	80.9	73.8
	PolarMask	89.8	80.4	73.5
	Ours	<b>91.8</b>	<b>82.3</b>	<b>75.6</b>
Back light	Mask R-CNN	85.3	78.1	69.3
	SOLO	88.6	79.4	71.8
	PolarMask	87.9	78.8	71.7
	Ours	<b>90.0</b>	<b>80.2</b>	<b>73.6</b>
Natural light	Mask R-CNN	86.8	79.6	71.8
	SOLO	88.9	80.6	72.3
	PolarMask	88.6	80.8	72.6
	Ours	<b>90.2</b>	<b>81.8</b>	<b>73.8</b>

value of each combined model, and obtained the experimental results shown in Table 3. The working mode of each module in the actual environment is discussed below.

As shown in Table 3, removing ICM will reduce the  $AP$  of the model by 1.3%. This shows that by adding the image correction module, the images under different lighting can be normalized to similar data distribution, which is equivalent to increasing the robustness of the model to complex environmental lighting. When MSEF-FPN is removed, the  $AP$  of the model decreases by 1.5%. This shows that adding MSEF-FPN can enhance the feature extraction ability of the image and reduce the interference of complex background to Apple detection results without increasing the amount of calculation.

**Table 3.** The results of Ablation experiments.

Method	ICM	MSEF-FPN	$AP$	$AP_{50}$	$AP_{75}$
Ours			85.3	80.1	74.0
Ours	✓		86.6	81.2	75.3
Ours		✓	86.8	82.0	76.0
Ours	✓	✓	<b>88.4</b>	<b>84.2</b>	<b>77.3</b>

## 5 Conclusion

In this paper, we propose a new instance segmentation method based on dual attention-guided network for Apple instance segmentation in complex environments, which solves the constraints of illumination, occlusion and overlapping

changes in environments, so as to realize the visual guidance of automatic picking. The CNN model with an image correction module and a instance segmentation module is constructed to meet the challenges of illumination, occlusion and overlap in complex environments. Experimental results show that the proposed algorithm performs better performance than the previous algorithms in instance segmentation. This enables the model to be deployed on the apple picking robot detector for automatic Apple detection.

## References

1. Saleem, M.H., Potgieter, J., Arif, K.M.: Automation in agriculture by machine and deep learning techniques: a review of recent developments. *Precis. Agric.* **22**(6), 2053–2091 (2021)
2. Maheswari, P., Raja, P., Apolo-Apolo, O.E., et al.: Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques-a review. *Front. Plant Sci.* **12**, 1247 (2021)
3. Bac, C.W., van Henten, E.J., Hemming, J., Edan, Y.: Harvesting robots for high-value crops: state-of-the-art review and challenges ahead. *J. Field Rob.* **31**(6), 888–911 (2014)
4. Jia, W., Tian, Y., Luo, R., Zhang, Z., Lian, J., Zheng, Y.: Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **172**, 105380 (2020)
5. Jia, W., Wang, Z., Zhang, Z., Yang, X., Hou, S., Zheng, Y.: A fast and efficient green apple object detection model based on Foveabox. *J. King Saud Univ. Comput. Inform. Sci.* (2022)
6. Patrício, D.I., Rieder, R.: Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. *Comput. Electron. Agric.* **153**, 69–81 (2018)
7. Chen, H., Sun, K., Tian, Z., et al.: BlendMask: top-down meets bottom-up for instance segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8573–8581 (2020)
8. Tang, Y., et al.: Recognition and localization methods for vision-based fruit picking robots: a review. *Front. Plant Sci.* **11** (2020)
9. Zhao, Z.-Q., Zheng, P., Xu, S.-T., Wu, X.: Object detection with deep learning: a review. *IEEE Trans. Neural Netw. Learn. Syst.* **30**(11), 3212–3232 (2019)
10. Wang, Z., Jia, W., Mou, S., et al.: KDC: a green apple segmentation method. *Spectrosc. Spectral Anal.* **41**(9), 2980–2988 (2021)
11. Vasconez, J.P., Delpiano, J., Vougioukas, S., Auat Cheein, F.: Comparison of convolutional neural networks in fruit detection and counting: a comprehensive evaluation. *Comput. Electron. Agric.* **173**, 105348 (2020)
12. Minaee, S., Boykov, Y.Y., Porikli, F., et al.: Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell.* <https://doi.org/10.1109/TPAMI.2021.3059968> (2021)
13. Lv, J., Wang, F., Xu, L., Ma, Z., Yang, B.: A segmentation method of bagged green apple image. *Sci. Hortic.* **246**, 411–417 (2019)
14. Sun, S., Jiang, M., He, D., Long, Y., Song, H.: Recognition of green apples in an orchard environment by combining the GrabCut model and Ncut algorithm. *Biosyst. Eng.* **187**, 201–213 (2019)

15. Ji, W., Gao, X., Xu, B.O., Chen, G.Y., Zhao, D.: Target recognition method of green pepper harvesting robot based on manifold ranking. *Comput. Electron. Agric.* **177**, 105663 (2020). <https://doi.org/10.1016/j.compag.2020.105663>
16. Linker, R., Cohen, O., Naor, A.: Determination of the number of green apples in RGB images recorded in orchards. *Comput. Electron. Agric.* **81**, 45–57 (2012)
17. Tian, Y., et al.: Fast recognition and location of target fruit based on depth information. *IEEE Access* **7**, 170553–170563 (2019)
18. Ahmad, M.T., Greenspan, M., Asif, M., et al.: Robust apple segmentation using fuzzy logic. In: 5th International Multi-Topic ICT Conference IEEE, pp. 1–5 (2018)
19. Bargoti, S., Underwood, J.P.: Image segmentation for fruit detection and yield estimation in apple orchards. *J. Field Rob.* **34**(6), 1039–1060 (2017)
20. Qi, C.R., Su, H., Mo, K., et al.: Pointnet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)
21. Sultana, F., Sufian, A., Dutta, P.: Evolution of image segmentation using deep convolutional neural network: a survey. *Knowl. Based Syst.* 106062 (2020)
22. Li, J., Liu, Z.M., Li, C., et al.: Improved artificial immune system algorithm for Type-2 fuzzy flexible job shop scheduling problem. *IEEE Trans. Fuzzy Syst.* (2020)
23. Jia, W., Zhang, Z., Shao, W., et al.: RS-Net: robust segmentation of green overlapped apples. *Precis. Agric.* (2021). <https://doi.org/10.1007/s11119-021-09846-3>
24. Anvari, F., Lakens, D.: Using anchor-based methods to determine the smallest effect size of interest. *J. Exp. Soc. Psychol.* **96**, 104159 (2021)
25. Liu, G., Nouaze, J.C., Touko Mbouembe, P.L., Kim, J.H.: YOLO-tomato: a robust algorithm for tomato detection based on YOLOv3. *Sensors* **20**(7), 2145 (2020). <https://doi.org/10.3390/s20072145>
26. Pan, X., Luo, P., Shi, J., Tang, X.: Two at once: enhancing learning and generalization capacities via ibn-net. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 464–479 (2018)
27. Wang, X., Kong, T., Shen, C., Jiang, Y., Li, L.: Solo: segmenting objects by locations. arXiv preprint [arXiv:1912.04488](https://arxiv.org/abs/1912.04488) (2019)
28. Xie, E., et al.: Polarmask: single shot instance segmentation with polar representation. arXiv preprint [arXiv:1909.13226](https://arxiv.org/abs/1909.13226) (2019)