Journal of King Saud University – Computer and Information Sciences 34 (2022) 7259–7270

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

DLNet: Accurate segmentation of green fruit in obscured environments

Jie Liu [a], Yanna Zhao [a], Weikuan Jia [a,*], Ze Ji [b]

[a] School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
[b] School of Engineering, Cardiff University, Cardiff CF24 3AA, United Kingdom

ARTICLE INFO

Article history:
Received 30 August 2021
Revised 22 September 2021
Accepted 27 September 2021
Available online 4 October 2021

Keywords:
DLNet
Obscured fruit
GAT
RS-RFP
Robust segmentation

ABSTRACT

To achieve more accurate recognition and segmentation of obscured fruit in natural orchard environments, DLNet model is proposed. The model is improved for the more challenging problem of segmenting overlapping fruit from homochromatic backgrounds without considering various damages. This approach is tantamount to construct the detection network RS-RFP and the segmentation network DLNet. RS-RFP extends Full Convolutional One-Stage Object Detection (FCOS). Specifically, Feature Pyramid Network (FPN) by adding Gaussian non-local attention mechanism to build Refined Pyramid Network (RFP) for refining semantic features generated continuously by Residual Network (ResNet) and FPN. The DLNet segmentation framework is composed of a dual-layer Graph Attention Networks (GAT) layer is constructed to model the image as two overlapping layers, where the top GAT layer detects the occluded object (occluded) and the bottom GAT layer infers the partially occluded instance (occlude). Display modeling of the two-layer structure occlusion relationship can naturally the boundaries between the occluded and occlude instances and consider their interactions. The experimental results show that the method outperforms earlier segmentation models and achieves metric values of 80.9% and 81.2% for Average Precision (AP) box and AP mask respectively. In a reasonable running time, it meets the requirements of accuracy and robustness for picking robots and provides a reference for segmentation of other fruits and vegetables.

1. Introduction

With the gradual maturation of deep learning, the transplantation of this new revolution to various industries for better results has become a common phenomenon, which has stimulated the development of autonomous robots in agriculture. Vision systems, as the most fundamental and important part of agricultural robots for resolving specified targets from complex and diverse scenes, which have been widely used in many practical applications, such as fruit yield estimation (Zhang et al., 2021), crop growth monitoring (Schima et al., 2016), and disease detection (Zhao et al., 2016). As for the visual recognition system, which is an important component of fruit and vegetable picking robots (Bauer et al., 2019), the accuracy, efficiency, and robustness of its fruit detection under complex background conditions will greatly affect the packing quality of picking robots. Therefore, a picking robot equipped with a stable vision recognition system will be the key to achieving efficient detection of target fruit is to realize intelligent management of orchards.

Machine learning plays a major role in image segmentation, and promising results have been achieved in green fruit segmentation. Arefid (Arefi et al., 2011) first removed the background in Red Green Blue (RGB) space, then combined RGB and Horizontal Situation Indicator (HSI) space to extract ripe tomato regions, and finally used shape features to locate fruit regions, and the overall accuracy of the algorithm was able to reach 96.36%. Dorj (Dorj et al., 2017) identified citrus with the help of color features and after a series of image processing methods to estimate fruit yield, but the color difference between citrus and leaves is obvious, there are few cases of mixed detection, which is relatively simple to identify. Tian (Tian et al., 2019) proposed a depth image-based target fruit localization method to fit the target region by locating the apple circle center and its radius through the depth image and its corresponding RGB spatial information respectively, but the method is difficult to locate the fruit circle center by depth image for the problems of fruit overlap and occlusion, and the robustness is poor

* Corresponding author.
E-mail address: jwk_1982@163.com (W. Jia).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2021.09.023
1319-1578/

in complex environment. Bhunia (Bhunia et al., 2020) proposed a novel feature descriptor to explore the relationship between hue (H) and saturation (S) channels in Hue-Saturation-Value (HSV) color space, combining color and texture information. Experiments show that the proposed descriptor is a significant improvement over existing descriptors for content-based color image retrieval. Bhattacharyya (Bhattacharyya et al., 2019) proposed a new method for obtaining specific gender information from facial images to extract information from frontal facial images to discriminate gender, which is still stable under the interference of background, illumination, intensity and facial expressions, thus improving the overall classification accuracy. These methods are often accompanied by a series of complex operations such as image preprocessing, feature selection and extraction, and the recognition effect of the model is easily involved in these operations. When the texture features on the fruit surface are missing due to light intensity, shape due to branch and leaf occlusion or overlap between fruits, and when the target fruit has the same color as the branches and leaves in the background and causes color interference, these problems can greatly reduce the recognition accuracy of such methods for the target fruit.

In recent years, with the development of deep learning and convolutional neural networks, the end-to-end detection process and the advantage of automatic extraction of image depth features, which eliminates many complex operations of traditional vision algorithms, it has attracted many researchers to apply them to target fruit localization and recognition. Bargoti (Bargoti and Underwood, 2017) first segmented apple images using a multistage perceptron and convolutional neural networks to extract apple targets in the image, and then used watershed segmentation and circular Hough transform method to identify and count apple targets. Jia (Jia et al., 2020) adapted Mask R-CNN (He et al., 2017) which is an instance segmentation model to apple target detection by improving the Residual Network (ResNet) (He et al., 2016) with Densely Connected Convolutional Networks (DenseNet) (Huang et al., 2017) as the feature extraction network of the new model to substantially improve the detection accuracy of apple targets in overlapping and branch-obscured environments. Chen (Chen et al., 2017) proposed a fully connected Convolutional Neural Network (CNN) based blob detector for extracting candidate regions in images, segmenting object regions, and using the subsequent of CNN counting algorithm to calculate the number of fruits. Gupta (Gupta et al., 2020) proposed a two-step method for Content-Based Image Retrieval (CBIR), registers image binary patterns and valley patterns and combines them with color histograms. This method overcomes the existing methods that use larger feature vectors and still have low detection accuracy. Ghose (Ghose et al., 2021) proposed a novel method for the recognition of ground terrain by modeling texture information to establish a balance between unordered texture information components and ordered spatial information to achieve an effective classification of ground terrain by a classifier. In the above deep learning-based detection model, its accuracy and applicability are significantly improved compared with traditional vision methods, but such methods require a large amount of computing and storage resources, and the speed is not yet able to meet the demand of real-time for picking robots.

Through the above mentioned domestic and international research status, in order to balance the relationship between accuracy and speed of target fruit segmentation and make the robot achieve the requirement of real-time operation in the heavily obscured orchard environment, this paper proposes DLNet instance segmentation model. The model consists of a detection network framework RS-RFP and a segmentation network framework DLNet, where RS-RFP is an extension of Full Convolution One-Stage Object Detection (FCOS) (Tian et al., 2019) by adding the non-local (Wang et al., 2018) attention block to the Feature Pyramid Network (FPN) (Lin et al., 2017) and constructing a Refinement Pyramid Network (RFP) to improve the accuracy of feature extraction. The segmentation of the severely occluded part of the DLNet network is achieved by constructing a two-layer graph attention network (GAT) (Veličković et al., 2017), the segmentation of the model under leaf and fruit occlusion interference can be realistically improved under the two-layer GAT structure, which satisfies the multiple requirements of speed, accuracy and robustness of each intelligent technique in practical applications. In general, this study has at least the following contributions:

(1) Embedding a non-local attention module and building a GAT structure to focus on information pixels while suppressing noise.
(2) The method set out in the present paper outperforms state-of-the-art models in terms of accuracy and robustness, and is more suitable for green fruit segmentation in complex scenes.
(3) Since DLNet eliminates the anchor frame, there is no need to reset the hypermastigote for a specific dataset, which means that the renamed model can be directly migrated to segmentation of other fruit.

The rest of this paper is organized as follows: Sect. 2 describes the image acquisition and processing and annotation of the associated datasets. Next, Sect. 3 introduces the detailed composition of the detection network RS-Net and the segmentation network DLNet and their improvements respectively. In Sect. 4 experiments verify that the method outperforms other methods in terms of precision, recall and robustness. Finally, the proposed method is summarized and the unresolved problems in this area are outlined, which are the future research directions.

2. Data collection and dataset creation

2.1. Data acquisition

To evaluate the segmentation effect of the model on green fruits, two datasets: unripe persimmon and green apple were collected and produced for the experiments in this paper, both of which were captured with the Sony Alpha 7 II camera, manually annotated with the target fruits in the images using labelme software, and uniformly transformed into MS COCO (Lin et al., 2014) dataset format, in order to adapt the model segmentation effect under the obscured environment, the COCO format dataset is further transformed into the bilayer annotation format required by this model for model learning. The persimmon dataset was collected from Shandong Normal University, Changqing District, Jinan City, Shandong Province, and the southern mountainous area of Jinan City, the persimmon dataset with 553 images and 2524 persimmon fruits labeled; the apple images were collected from the apple production base in Fushan District, Yantai City, Shandong Province, with 268 images and 649 apple fruits labeled. Both datasets contain images of fruit captured in various environments such as different time periods, different weather, different light angles, and different shading conditions. Taking the persimmon dataset as an example, Fig. 1 shows some of the actual images under different situations respectively; Table 1 shows the acquisition time period, image and fruit distribution, and training/validation set division of the persimmon dataset.

2.2. Dataset production and dataset enhancement

The dataset was labeled with labelme software, and the resolution was uniformly reduced to 600 × 400 pixels before labeling.

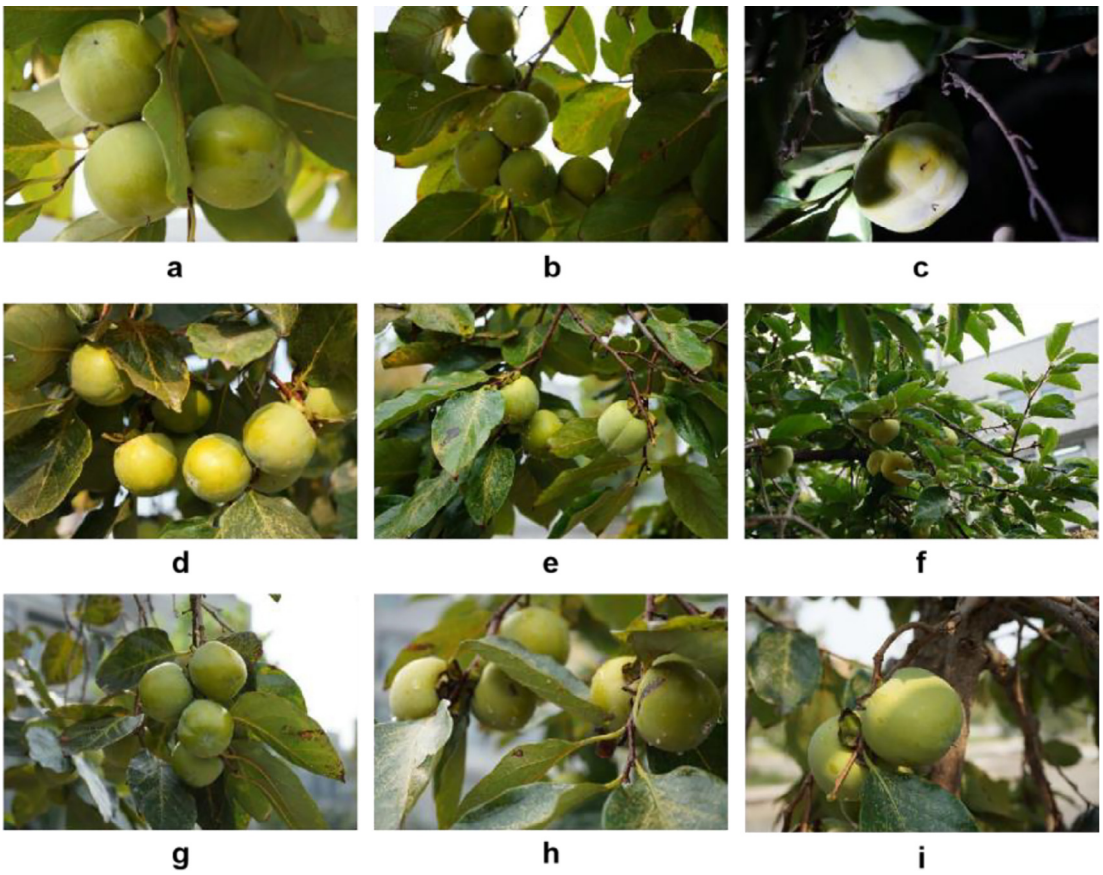


Fig. 1. Actual persimmon images under complex natural scene. Note: a-c show images at different light intensities; d-f show different types of shading (fruit overlap, leaf shading, branch shading); g-i show images at different light angles.

Table 1
Distribution of persimmon dataset.

Time Scenes Divide	Training set Pictures (sheets)/ Fruits (pieces)	Validation set Pictures (sheets)/ Fruits (pieces)	Sum Pictures (sheets)/ Fruits (pieces)
Morning	65/266	28/84	93/350
Noon	87/278	37/142	124/420
Afternoon	69/521	29/129	98/650
Evening	87/336	38/125	125/461
After the rain	80/464	33/179	113/643
Sum	388/1865	165/659	553/2524

Note: Table of image acquisition and data set division. The acquisition was performed under natural sunlight during the daytime, and LED lights were used to assist in illumination at night. For the images acquired under each time period, the training set and the validation set are divided in a ratio of 3:1 for the number of images.

The minimum external matrix of each fruit in the labeled image was used as the true frame and the corresponding json file was generated, and it was randomly combined into training set and dataset according to 7:3, where the number of images in the training set of persimmon dataset was 398 and the number of images in the test set was 170 images. Since the number of images and labeled fruit in the apple dataset was small, the images after the labeling were enhanced randomly. The enhancement types included brightness enhancement, contrast reduction, fogging, gaussian noise, impulse noise, Poisson noise. As showed in Fig. 2, each enhancement type was divided into different enhancement degrees, and a total of 5290 images were finally generated. The enhanced images generated from the original images share the annotation information in the same json file, and the training set and validation set are divided in the ratio of 7:3 in each interference degree of each interference type. Finally, a total of 3703 images are obtained from the apple training set and 1587 images from the validation set, and the annotation files in MS COCO dataset format are generated respectively.

3. DLNet double-layer occlusion segmentation model

To improve the accuracy and efficiency of segmentation of green fruit in an orchard shading environment, an accurate and efficient DLNet segmentation model is proposed in this paper. The framework of the new model is presented in Fig. 3 below, which consists of three parts: (1) feature extraction; (2) feature refinement; (3) result prediction. First, "feature extraction" and "feature refinement" phases consist of the detection network RS-RFP, which consists of three steps: extraction, fusion, and refinement by ResNet, FPN and RFP respectively (see Fig. 4 for details). The segmentation network DLNet, which uses a two-layer GAT structure, in which the top GAT detects occluded objects and the bottom GAT infers partially occluded instances. With the two-layer GAT structure, the instances of the occluded part are fetched and the mask is generated by Fully Convolutional Networks (FCN) (Long et al., 2015) to generate the detailed region where the fruit is located.

3.1. Feature extraction RS-RFP detection network

The feature extraction RS-RFP detection network consists of three parts: extraction, fusion and refinement, which are handled

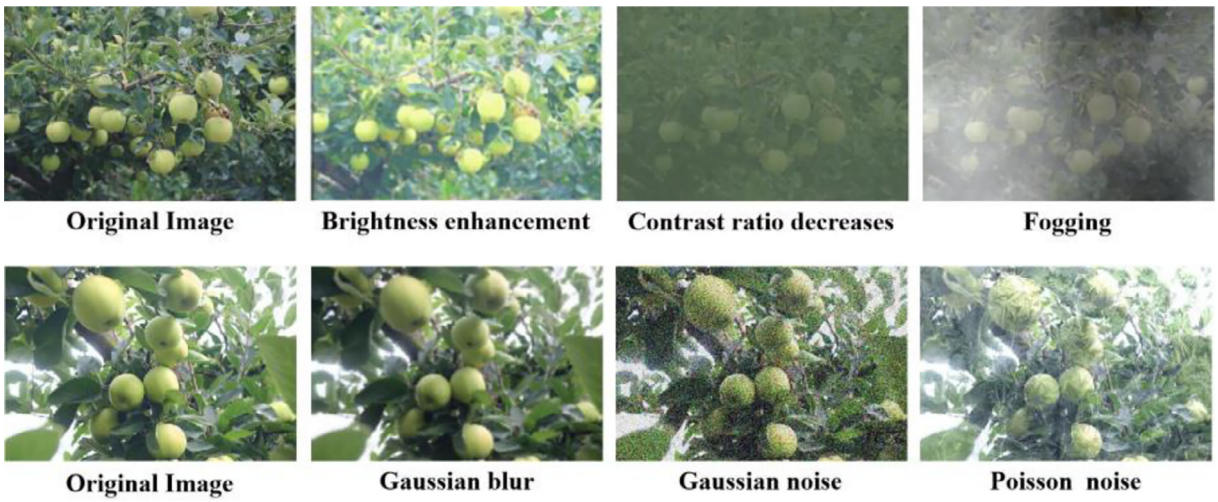


Fig. 2. Examples of different types of apple image enhancement.

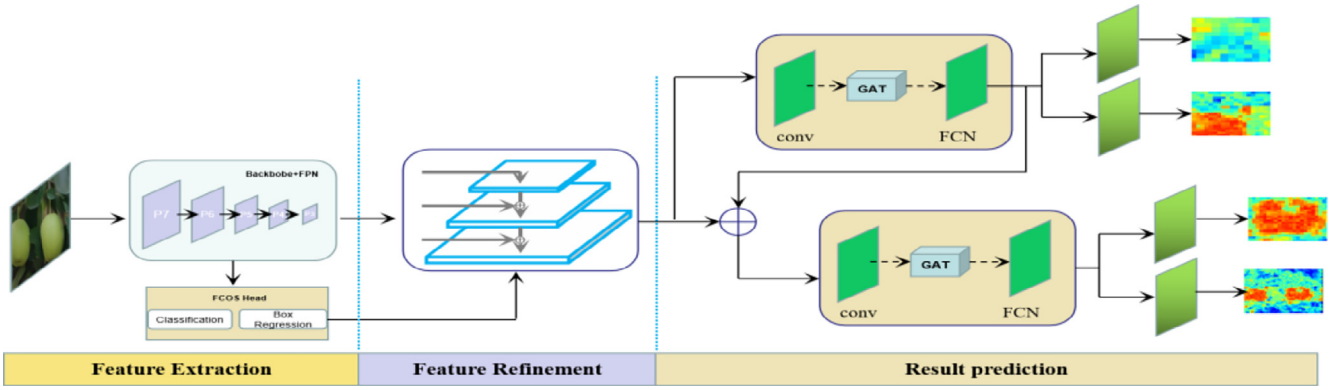


Fig. 3. Flow chart of DLNet.

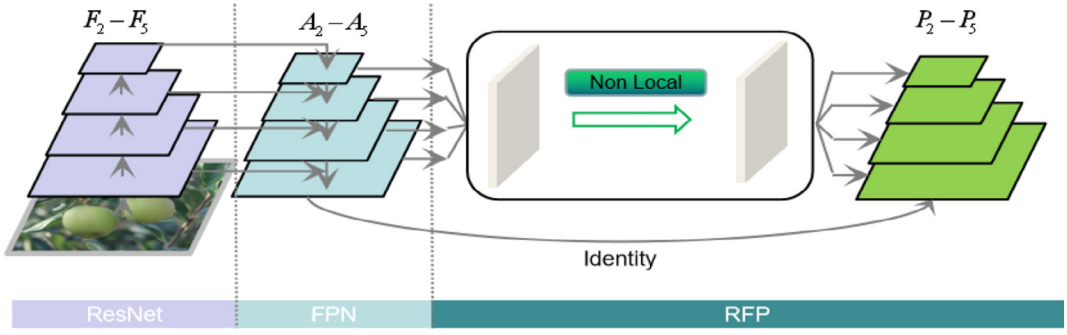


Fig. 4. Flow chart of RS-RFP. Note: The detection network contains a total of three parts, the backbone network ResNet and FPN to extract the image features, and RFP to refine the extracted features, a detailed diagram of the attention module embedded in RFP is shown in Fig. 5.

by ResNet101, FPN and RFP respectively. The combination of ResNet and FPN can lead to gradient disappearance and explosion as the depth of the network increases, which lead to degradation of the model. Therefore, based on the efficient feature extraction capability of ResNet and FPN, RFP is introduced, thus effectively solving this paradoxical phenomenon and improving the discriminative ability of deeper networks.

3.1.1. ResNet + FPN

In general, deep high-level features in the ResNet101 backbone network have more semantic information, while the shallow low-level features are more descriptive in content. Although the final output feature map of the ResNet101 network contains rich semantic information, after continuous down sampling operations (convolution and pooling), it will make its resolution very low and detail information such as boundary is basically lost, which will make the semantic information of smaller objects severely diluted and eventually lead to detection failure, so its extracted feature values are suitable for predicting large scale targets. Considering the distance between the robot and the object, and the small area of the obscured object, the vision system design of the picking robot also needs to accurately identify smaller areas of fruit in the image, so FPN is introduced in this model architecture.

In this study, final residual blocks conv2, conv3, conv4, conv5 of ResNet101 are taken, whose output feature maps are $\{A_2, A_3, A_4, A_5\}$, and its feature maps are fused according to top-down and lateral connections to obtain the $\{F_2, F_3, F_4, F_5\}$. FPN mainly solves the target multi-scale prediction problem in detection, which constructs a feature pyramid by fusing the semantic information of the deep feature map and the detail information of the shallow feature map, and distributes the targets to be detected at different scales to the feature maps at different levels in the pyramid responsible for prediction.

3.1.2. RFP

The features extracted by ResNet and FPN can be used as the basis for detection, and great progress has been made and high accuracy can be achieved, but the application of ResNet + FPN network for fruit detection in complex orchard environment will have the following problems. On the one hand, fruit images are collected in the complex orchard environment. These images are affected by unfavorable factors such as illumination, overlap and especially occlusion, making the fruit area in the captured image incomplete. On the other hand, the integrated feature extraction method should have balanced information from the semantic features of each pixel of each image, but in the ResNet + FPN structure will make the integrated features pay more attention to the semantic information of adjacent pixels and less attention to other resolutions, and the semantic information contained in non-adjacent levels will be diluted in each fusion during the information flow. Therefore, in order to solve the above two dilemmas, RS-RFP network is added RFP which is embed non-local modules on top of FPN to obtain and refine more semantic feature information, whose structure is shown in Fig. 4 above, and the specific implementation details are shown in Fig. 5 below.

In this paper, in order to set the non-local block more efficient by adding a maximum pooling layer behind φ and ϕ in Fig. 5, the number of channels of w_θ, w_ϕ and w_ω would set to half of the number of channels of $\times$, thus forming a bottleneck that will be able to reduce the computation by half. w_z then re-approach to the number of channels of $\times$ to ensure that the input and output dimensions are consistent. After using the down sampling operation, the outputy$_i$ becomes the following equation:

$$y_i = \frac{1}{c(\hat{x})} \sum \forall_j f(x_i, \hat{x}_j) g(\hat{x}_j) \tag{1}$$

where x denotes the input; $f(x_i, x_j)$ is used to calculate the pairwise relationship between I and all possible associated positions j; $g(x_j)$ is used to calculate the eigenvalues of the input signal at position j; and $c(x)$ is the normalization parameter.

3.2. DLNet segmentation model

In images in heavily occluded environments, multiple overlapping objects in the same bounding box during segmentation may lead to confusion between instance profiles from real objects and occluded boundaries. For example, the masked header design of Mask scoring RCNN (Huang et al., 2019) directly regresses masking with a fully convolutional network that ignores the overlapping relationship between masked instances and objects. To alleviate this limitation, DLNet extends the existing two-stage instance segmentation approach by adding a two-layer GAT structure to the traditional target prediction pipeline, so that the interactions between objects in the region of interest and thus the specific real objects and masks can be well considered in the mask regression stage.

3.2.1. Double-layer GAT structure

In recent years, Graph Convolutional Network (GCN) (Kipf and Welling, 2016) has been used to model long-term relationships in images and videos, and highly overlapping targets, closes can segment pixels belonging to the same part of the occluded object into disjoint sub-regions. However, since GCN assumes that graphs are directed and cannot handle dynamic graphs, and cannot assign different weight to each neighboring point, GAT is introduced. Based on the non-local properties of GAT, DLNet is adopt GAT as the basic block, where each graph node represents a single pixel on the feature map. To explicitly model closed regions, the model extends the single GAT block into a two-layer GAT structure as showed in Fig. 3, constructing two orthogonal graphs under a single generic framework.

In this paper, the segmentation part of the model is designed simply and efficiently, consisting of a 3×3conv, followed by a GAT layer and an FCN layer, after which the output is fed to the up sampling and 1×1convolution layers to obtain a channel feature mapping for joint boundary and mask prediction. Implementation of GAT is performed through the Dual Attention Network (DANet) (Fu et al., 2019a) modules, which are divided into a position attention module and channel attention module, the built GAT structure is shown in Fig. 6 below.

3.2.2. DLNet workflow

In the GAT structure in this, an adjacency graph $G = \langle V, \varepsilon \rangle$ needs to be given, where there is edges εbetween nodes V. The graph convolution operation is represented as:

$$O = \sigma(AXW_g) + x \tag{2}$$

where $X \in R^{N \times K}$ is the input feature map; $N = H \times W$ is the number of pixel grids within the ROI region; K is the feature dimension of each node; $A \in R^{N \times N}$ is the functional similarity that defines the

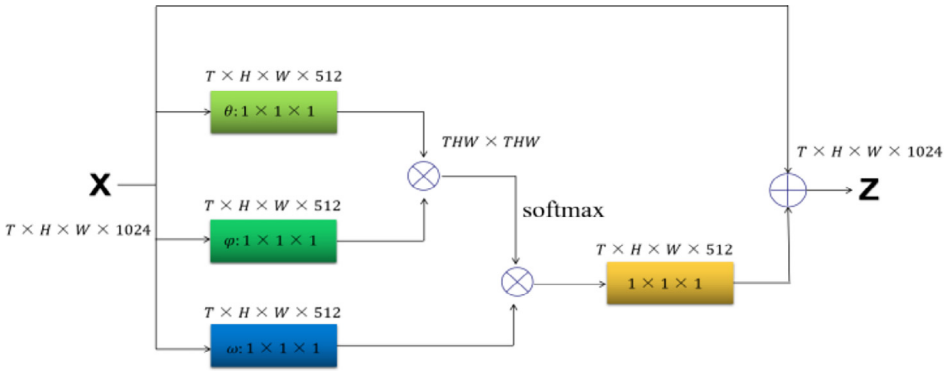


Fig. 5. Detailed description of attention module which illustrated in RFP section of Fig. 4.

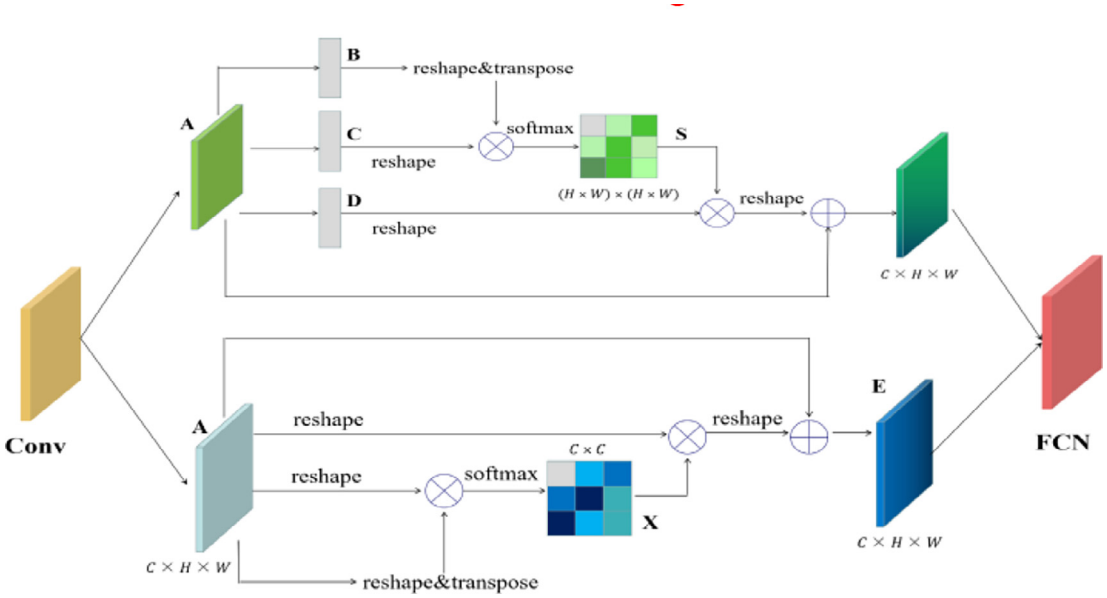


Fig. 6. The built GAT structure.

nodes of the adjacency matrix graph; $W_g \in R^{K\times K}$ is the output transformation matrix, in this model $K' = K$; and the output features $O \in R^{K\times K}$consist of node features updated by global information propagation within the entire graph layer, which is normalized and Rectified Linear Unit (ReLU) (Krizhevsky et al., 2012) function of the nonlinear function $\sigma(.)$is obtained. In the two-layer GAT structure, it is further defined g^ias the with graph, X_{roi}as the input ROI features, and W_f as the weight of the FCN layer, so that the complete formula is:

$$O^1 = \sigma\{A^1[\sigma(A^0 X_{roi} W_g^0) + X_{roi}]W_g^1\} + O_f \tag{3}$$

$$O_f = \sigma(A^0 X_{roi} W_g^0) W_f^0 + X_{roi} \tag{4}$$

To connect the two GAT blocks, the output feature $A^0$of the first GAT is added directly to X_{roi} to obtain the fused occlusion-aware feature O_f, $O_f$which is the input of the second GAT layer, and the output O^1 is used for occlusion mask prediction.

In the process of segmentation by the two-layer GAT structure, the feature information obtained in the RS-RFP detection network is processed, especially for the discrimination of the occluded fruit, and the occluded part is processed in steps. In the first choice, occluded part and the obscured part are distinguished and then sent to the respective processing GAT layers, and finally the information is integrated and processed to output the final predicted image. Among them, the first GAT layer is used to detect contours and process occluded instances to achieve contour prediction and mask regression for occluded instances, and the second GAT layer is used to process occlude instances to achieve contour prediction and mask regression for occlude instancess. Based on the attention mechanism in GAT, it can focus more on functional information and reduce noise interference. After such a two-layer GAT processing, the occluded and occluded are processed separately and then integrated to achieve accurate segmentation for the mask as showed in Fig. 7 below.

The bilayer GAT structure constructs a new semantic graph space for the enclosed region additionally compared to the previous single-layer structure of the class of unknown mask headers, which has only binary labels (foreground/ background) per pixel. The model explicitly distinguishes the work of the two-layer occlusion structure, and the overlap between the two layers can be directly identified as the occlusion boundary, so that it can be distinguished from the real object contours.

3.3. Loss function

One of the important factors determining the effectiveness of the model for fruit segmentation is the design of the loss function. Based on the prediction objectives of each branch, the task type, the proportion of positive and negative samples, the loss function showed below is used for iterative optimization of the model.

According to the structural analysis of the model, the loss function of the model should be composed of three parts: the loss generated in the detection phase, the loss of the occluded branch and the loss of the occlude branch. The overall loss function equation is shown as following.

$$L = L_{Detect} + L_{Occluder} + L_{Occludee} \tag{5}$$

Regarding the loss generated by the model in the detection phaseL_{Detect}, it is further composed of the losses generated by the three branches of Classification, Regression, and Centerness. Since a picture in which the target fruit occupies a relatively small area compared with the background and undergoes a factor shrinkageσ, there is an imbalance problem between positive and negative samples in the training phase. In order to take into account the above disadvantages and simplify the calculation, so Classification, Regression, and Centerness branches are chosen to be calculated by Focal Loss (Lin et al., 2017), IoU (intersection of union) Loss (Yu et al., 2016), and BCE Loss (de Boer et al., 2005) respectively, and the overall loss of the detection part of the model function is shown below:

$$L_{Detect} = L(\{p_{x,y}\}, \{d_{x,y}\}, \{center_{x,y}\}) = L_{cls} + L_{regression} + L_{centerness} \tag{6}$$

$$L_{cls} = \frac{1}{N_{pos}} \sum_{x,y} L_{class}(p_{x,y}, p^*_{x,y}) \tag{7}$$

$$L_{regression} = \frac{\lambda}{N_{pos}} \sum_{x,y} L_{regression}(d_{x,y}, d^*_{x,y}) \tag{8}$$

$$L_{centerness} = \frac{\beta}{N_{pos}} \sum_{x,y} L_{centerness}(center_{x,y}, center^*_{x,y}) \tag{9}$$

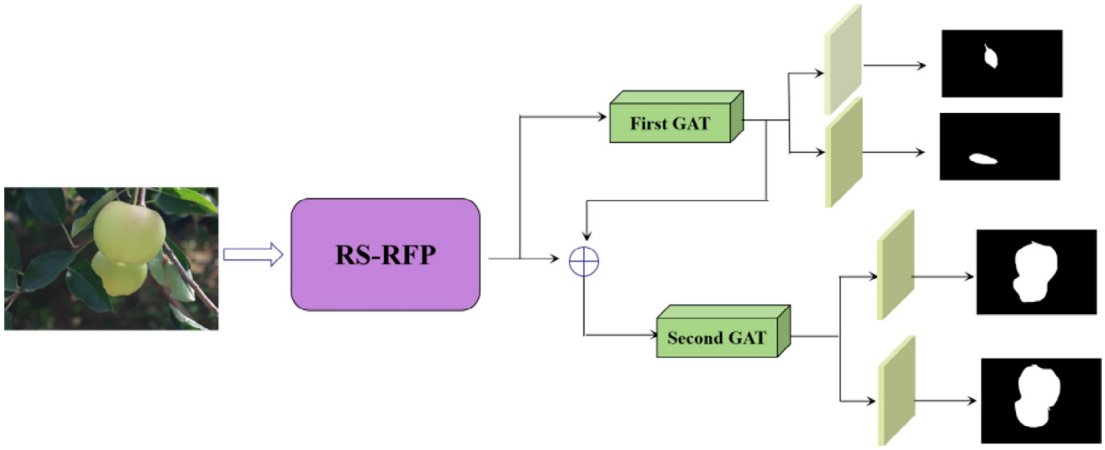


Fig. 7. The workflow diagram of GAT. Note: The first layer of GAT will extract features for segmentation of the occluded part, and the occluded part will be found and segmented; the second layer of GAT will segment the occluded object, and then the results obtained from the first layer of GAT will be merged to generate the final segmented image.

In the above equation, $p_{x,y}, d_{x,y}, center_{x,y}$ is the prediction value of classification branch, regression branch, and centrality branch at spatial location (x,y) respectively, $p^*_{x,y}, d^*_{x,y}, center^*_{x,y}$ corresponds to the training target at spatial location (x,y), and among the three loss terms, $L_{regssion}, L_{centerness}$is only for positive samples, N_{pos}denotes the number of positive samples, and λ, β is the balance coefficient of each loss term.

The model produces loss functions of $L_{Occludee}$and $L_{Occluder}$ in the occlude branch and occluded branch of the partitioned network with the following functional formulas as shown in Eq. (10) and Eq. (11).

$$L_{Occludee} = \lambda_1 L_{Occ-B} + \lambda_2 L_{Occ-S} \tag{10}$$

$$L_{Occluder} = \lambda_3 L'_{Occ-B} + \lambda_4 L'_{Occ-S} \tag{11}$$

Regarding the classification lossL'_{Occ-B} of boundary detection for segmented occludeds in Eq. (12), the following equation is shown.

$$L'_{Occ-B} = L_{BCE}(W_B F_{occ}(X_{rot}), gT_B) \tag{12}$$

where L_{BCE} denotes the binary cross-entropy loss, expressed as following.

$$L_{BCE}(x, class) = weight[class]\left(-x[class] + \log\left(\sum_j \exp(x[j])\right)\right) \tag{13}$$

F_{occ}denotes the nonlinear transformation function of the occlusion modeling module; W_B is the weight of the boundary predictor; X_{roi} is the shear FPN feature map given by the Roi Align operation of the target region; and gT_B is the ready-made encloser boundary, which can be easily calculated from the mask annotation.

With respect to the classification loss L'_{Occ-s} in Eq. (12) for modeling the occluded of the segmented occluded, the following equation is shown.

$$L'_{Occ-s} = L_{BCE}(W_S F_{occ}(X_{roi}), gT_S) \tag{14}$$

where $F_{occ}(X_{roi})$ is for the shared features of the joint optimization using boundary prediction in the mask prediction of the mask; W_S denotes the trainable weight of the predicted values of the 1×1 convolutional layer segmentation mask; gT_s denotes the mask labeling of the mask. Above $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ is the hypernatremia weight for balancing the loss function, which are tuned to $\{0.5, 0.25, 0.5, 1.0\}$ respectively on the validation set.

4. Results and analysis

In check to see the effectiveness of DLNet model for green fruit recognition, the following experiments are conducted and the results are analyzed. The experiments were firstly conducted by using both per-training and direct training to obtain the models and compare their effects on the experimental results. Then, the optimal training model is selected, evaluated on the validation set of both persimmon and apple fruit and the experimental data are analyzed. Finally, the state-of-the-art algorithms for each type of target detection and segmentation are selected to test and compare the difference in detection and segmentation performance of the models when segmenting green fruit.

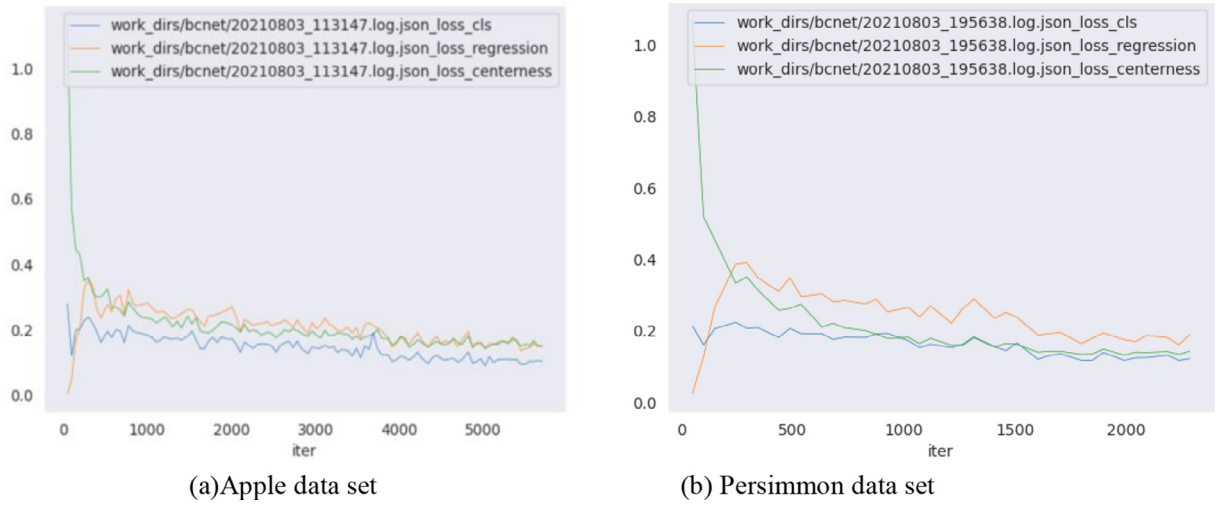
4.1. Experimental implementation details

All relevant experiments involved in this paper were done on the same server device with the main configuration environment of Ubuntu 16.04 OS, 32 GB Tesla V100 graphics card and 10.0 CUDA environment. All models were built using the Python language and the Pytorch 1.4 deep learning library with the help of relevant modules in the Detectron framework.

4.1.1. Training phase

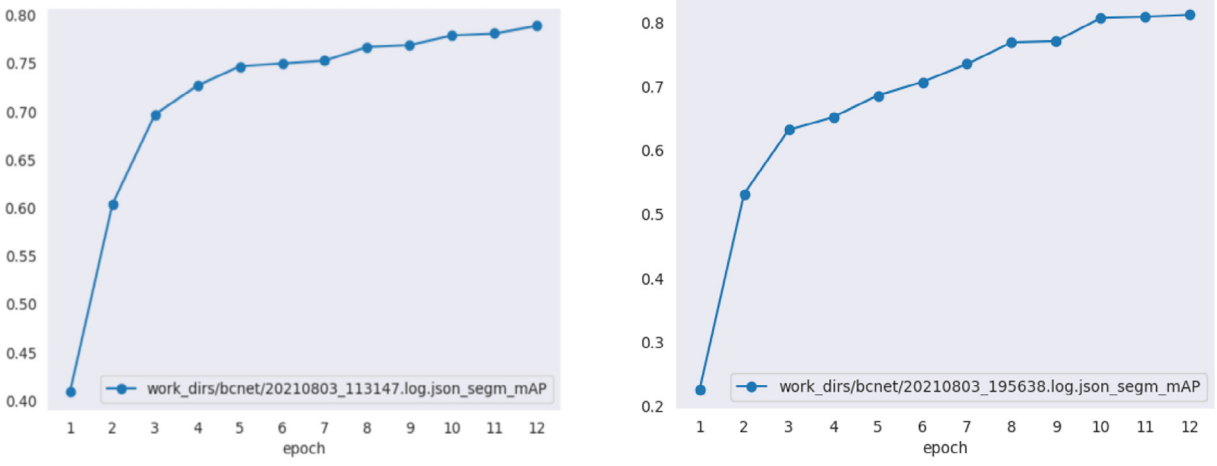
Before the formal training, 1586 apple images are used for per-training. The parameters after the per-training were migrated to the DLNet network as initialization parameters to better improve the accuracy and robustness of the model.

For formal training, the mini-batch was utilized to iteratively train 12 epochs, using 2 samples per iteration as a batch. The loss changes are generated by the three branches during training are shown in Fig. 8, with the horizontal axis showing the number of iterations and the vertical axis showing the loss values. After each training epoch was evaluated on the validation set, the obtained segmentation Average Precision (AP) change graph is given in Fig. 9. ResNet101 was used as the base network to extract the image features, and Batch Normalization (BN) (Ioffe and Szegedy, 2015) was used for regularization when the weights were updated; when constructing the FPN, a 5-layer pyramid hierarchy $\{p_l\}$ (l = 3,4,. . . ,7) fused features, the number of channels per layer is 256, and the down sampling multiplier is 2^lrespectively; the shrinkage factor σ is set to 0.4, the mapping region on the target fruit frame corresponding to the feature map is shrunk 0.4 times as the positive sampling region; the weights are regularized using



(a)Apple data set (b) Persimmon data set

Fig. 8. Loss function variation curves of the two datasets in the training phase.



(a)Apple data set (b) Persimmon data set

Fig. 9. Plot of mAP variation in the training phase for the two datasets.

BN each time they are updated, and the model parameters are updated using Stochastic Gradient Descent (SGD) (Bottou, 2010), with the learning rate, weight decay, and momentum set to 0.0025, 0.0001, and 0.9 respectively. The input network is uniformly sized to (1200,800) before training, and is sequentially reprocessed with random flipping, regularization, and padding operations.

4.1.2. Testing phase

The images are also per-processed with cropping, random flipping, regularization and padding before in putting into the network; after the network inference, the low quality prediction frames with confidence less than 0.05 are excluded first, and then the prediction frames with too much overlap are screened with NMS, using IoU equal to 0.5 as the threshold, and after the screening is completed, the prediction frames with at most the top 100 confidence are retained for each image in order of confidence, top 100 confidence frames are retained for each image.

4.2. Evaluation indicators

In this experiment, IoU = 0.5 between the model prediction box and the labeled box is used as the threshold to classify them as belonging to True Positive (TP) or False Positive (FP), count their number and calculate Precision and Recall according to the formula.

$$\text{Precision} = \frac{\text{TPs}}{\text{TPs} + \text{FPs}} \tag{15}$$

$$\text{Recall} = \frac{\text{TPs}}{\text{TPs} + \text{FNs}} \tag{16}$$

In Eq. (15) and Eq. (16), TPs, FPs, and FNs denote the number of true positive samples, the number of false positive samples, and the number of false negative samples under the specified confidence level and for IoU threshold, respectively. Finally, the AP index is used to objectively and comprehensively judge the model detection effect, as showed in Eq. (17).

$$AP_{IoU=i} = \frac{1}{101} \sum_{r \in R} \max_{\tilde{r}:\tilde{r} \geq r} p\left(\tilde{r}\right) \tag{17}$$

In Eq. (17), R ∈ [0, 0.01, 0.02, . , 1], r represents the value taken as the recall rate, and p is the accuracy rate corresponding to the value taken as the recall rate. Through the above equations, accuracy and recall are evaluated together to obtain the approximate

value AP (default i = 0.5) of Area Under Curve (AUC) under the specified IoU threshold and used as the evaluation index for the following experiments.

4.3. Model segmentation effect

After the training of the network is completed, and the optimal model is selected after performance evaluation, the model uses different IoU thresholds and AP and Average Recall (AR) values under different target fruit scale ranges as the evaluation index of the model to assess the overall performance of the network. In addition, several images containing persimmon and apple fruit under mixed interference conditions are selected such as overlapping, branch and leaf occlusion, nighttime, distant view, after rain and backlight for segmentation, studied and analyzed the segmentation effect maps. The segmentation effect of DLNet for both fruits is shown in Fig. 10, where persimmon fruits are shown on the left and apple fruits are shown on the right. The detailed evaluation results of the model on the two fruit validation sets are shown in Table 2.

where AP^{b}_{50},AP^{s}_{50} are the AP values of the model for the border and the AP values of the mask under the threshold of IoU = 0.5; mAP^{b}, mAP^{s} are the AP values of the model for the predicted border and the predicted mask of the network under the threshold of [0.5, 0.55, 0.6, . . . , 0.95] and averaging the 10 AP values obtained; furthermore, mAP_S, mAP_M, and mAP_L are the combined evaluation results of the model for small-scale fruit, medium-scale fruit, and large-scale fruit prediction results in the three scale ranges [0,322], [322,962], and [962,INF] respectively.

Although DLNet has been able to recognize most of the shaded fruits for the shading situation in the orchard, there are still cases where the light, leaves and fruits are too close to each other in color, and the fruits are not recognized due to severe shading, as shown in Fig. 11 below.

4.4. Algorithm comparison

To further illustrate the effectiveness of the model for target fruit segmentation, the performance of current advanced target detection and instance segmentation algorithms were run and evaluated on the same dataset and the same configuration of the experimental platform from the perspective of both detection and segmentation performance, and the differences in recognition performance between them and DLNet were compared. The finally experimental conclusions were drawn, as shown in Table 3.

As shown in Table 3 above, different types of target detection and instance segmentation algorithms are selected for comparison with DLNet, including anchor-frame-based two-stage algorithms: Faster R-CNN (Ren et al., 2015), Mask R-CNN, MS R-CNN; anchor-frame-based single-stage algorithms: SSD512 (Liu et al., 2016), YOLO v3 (Redmon and Farhadi, 2018), YOLACT (Bolya et al., 2019a), YOLACT++ (Bolya et al., 2019b), SOLO (Wang et al., 2020) and single-stage algorithms without anchor frames: FCOS, PolarMask (Xie et al., 2020), RetinaMask (Fu et al., 2019b), BCNet (Ke et al., 2021), "-" indicates that the model does not have the ability to predict borders or masks.

As showed in Table 3 above, compared with other algorithms, DLNet has the highest comprehensive evaluation index mAP and mAR values in terms of detection and segmentation accuracy, 80.9% and 81.2% respectively, which are higher than other three different types of algorithms. In addition to considering the accuracy of segmentation, it also need to consider the segmentation speed of the algorithm in recognizing an image on GPU on average. The model needs to reduce the segmentation time while ensuring the accuracy, and it is hard to really put it into use if the segmentation time does not reach the requirement of real-time. As showed in Fig. 12 below, the time required for the segmentation algorithm in Table 3 above to segment an image on average on the same dataset is listed.

The above analysis shows that the DLNet model can achieve higher detection accuracy with simpler model structure and less computation, and can achieve high efficiency in speed and accuracy at the same time and adapt to complex orchard environment, which can ensure more stable and efficient operation quality with less power consumption when deployed to mobile picking equipment.

4.5. Validation on the COCO dataset

To further validate the performance of the network, publicly available standard datasets for testing on the network was used. The COCO2014 dataset was selected, and four latest segmentation algorithms, Mask R-CNN, MS R-CNN, RetinaMask, and YOLACT were selected to compare the accuracy with the DLNet model on

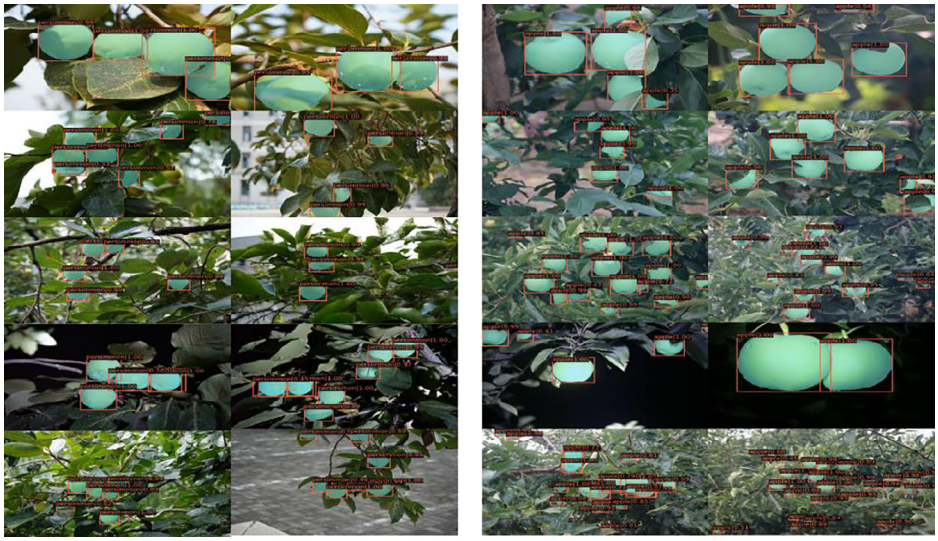


Fig. 10. Segmentation effects of the model in different interference scenes.

Table 2
Evaluation results of DLNet network on two validation sets.

Persimmon				Apple			
Bbox		Segm		Bbox		Segm	
Metric	Value	Metric	Value	Metric	Value	Metric	Value
mAP^b	80.9%	mAP^s* MERGEFORMAT	81.2%	mAP^b* MERGEFORMAT	82.8%	mAP^s* MERGEFORMAT	78.9%
AP^b_{50}	90.3%	AP^s_{50}* MERGEFORMAT	89%	AP^b_{50}* MERGEFORMAT	86.4%	AP^s_{50}* MERGEFORMAT	84.8%
mAP^b_s	44%	mAP^s_s* MERGEFORMAT	42.2%	mAP^b_s* MERGEFORMAT	45.7%	mAP^s_s* MERGEFORMAT	44.8%
mAP^b_m	73.6%	mAP^s_m* MERGEFORMAT	74%	mAP^b_m* MERGEFORMAT	70.6%	mAP^s_m* MERGEFORMAT	66.5%
mAP^b_l	85.6%	mAP^s_l* MERGEFORMAT	86.8%	mAP^b_l* MERGEFORMAT	88.7%	mAP^s_l* MERGEFORMAT	87.7%
mAR^b	80.4%	mAR^s* MERGEFORMAT	81%	mAR^b* MERGEFORMAT	84.1%	mAR^s* MERGEFORMAT	80.1%
mAR^b_s	45.9%	mAR^s_s* MERGEFORMAT	47.2%	mAR^b_s* MERGEFORMAT	46.9%	mAR^s_s* MERGEFORMAT	43.2%
mAR^b_m	80.1%	mAR^s_m* MERGEFORMAT	80.8%	mAR^b_m* MERGEFORMAT	82.8%	mAR^s_m* MERGEFORMAT	78.9%
mAR^b_l	89.8%	mAR^s_l* MERGEFORMAT	90.7%	mAR^b_l* MERGEFORMAT	91.6%	mAR^s_l* MERGEFORMAT	89.9%

Fig. 11. Image with missed segmentation. Note: (a) Leaves are mixed with fruits; (b) leaves are heavily obscured; (c) lighting is too dim.

Table 3
Recognition results of each model on both datasets.

Methods	Persimmon Dataset		Apple Dataset	
	mAP^b	mAP^s	mAP^b	mAP^s
two-stage anchor-based				
Faster R-CNN	72.3	–	82.4	–
Mask R-CNN	71.8	72.3	81.5	74.3
MS R-CNN	72.9	72.5	80.6	76.2
one-stage anchor-based				
SSD512	64.1	–	75.1	–
YOLO v3	69.6	–	82.2	–
YOLACT	58.0	61.0	67.4	75.6
YOLACT++	70.2	69.1	78.0	78.8
SOLO	–	58.6	–	76.4
one-stage anchor-free				
FCOS	68.8	–	81.4	–
PolarMask	57.7	54.6	69.9	68.7
RetinaMask	72.4	71.6	81.8	73.6
BCNet	76.2	75.4	80.1	77.8
ours				
DLNet	80.9	81.2	82.8	78.9

the COCO dataset, and their comparison results are shown in Table 4 below.

According to Table 4 above, DLNet algorithm is significantly higher than the other four algorithms in terms of accuracy, indicating a good segmentation performance, and is second only to the YOLACT algorithm in terms of time, but both in terms of model capacity and segmentation accuracy, DLNet makes up for its small loss in speed with a simpler architecture, better computation and higher segmentation accuracy. Therefore, after the above analysis, this method achieves good results in terms of segmentation accuracy and time, with strong generalization ability and robustness.

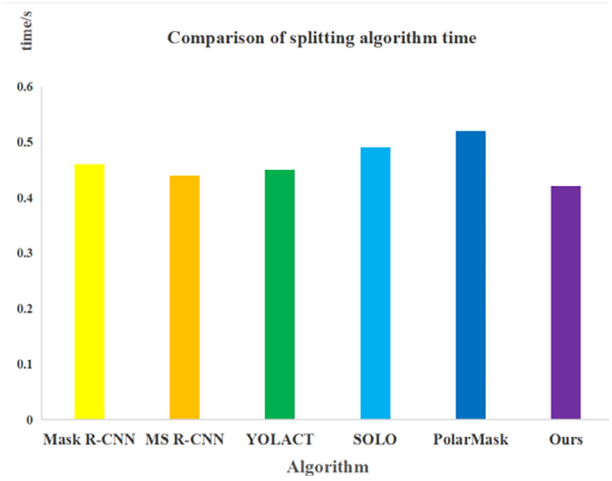


Fig. 12. Time comparison of different segmentation algorithms.

Table 4
Comparison results of different algorithms on coco dataset.

Method	Backbone	AP_S	AP_{50}	AP_{75}	Time
Mask R-CNN	R-101-FPN	35.7	58.0	37.8	116.3
MS R-CNN	R-101-FPN	38.3	58.8	41.5	116.3
RetinaMask	R-101-FPN	34.7	55.4	36.9	166.7
YOLACT	R-101-FPN	31.2	50.6	32.8	42.7
DLNet	R-101-FPN-RFP	40.4	60.1	42.3	53.2

5. Conclusion

In the unstructured orchard environment, for the segmentation challenges in the occluded environment, this study takes green fruit as the research object and proposes DLNet, which is a green

target fruit segmentation model in the occluded environment. The model consists of two parts, the first part is the detection network RS-RFP, which extends the FCOS generation by adding the embedded Gaussian attention module. Based on ResNet and FPN, the newly proposed RFP is added so that similar semantic features achieve mutual gain and reduce the influence of adverse factors such as occlusion, illumination, and overlap. The second part is the DLNet segmentation network, which uses GAT as the basic module, depended on its non-local properties as well as the attention network. To explicitly model occluded regions, the single GAT block is extended to a two-layer GAT structure to decouple the overlap relationship. In which the first GAT layer is used for occlusion prediction and the second GAT layer performs occluded modeling, which is used to guide the target (occluded) object segmentation through the rich auxiliary prediction information provided by the first GAT layer, such as shape and position prediction. The experimental results show that the new method is highly accurate in detecting and segmenting green target fruit, and robust under various interference conditions. Report to segmentation algorithms such as YOLACT, the model consumes less computation and storage, has a more concise architecture design, and is faster in segmentation. In the case of leaf occlusion, similar color to the background, overlapping and various lighting effects, the DLNet model can aggregate the green fruit information in the whole images during segmentation detection and suppress the background interference noise to achieve better segmentation results with minimal computational resources.

The new model achieves efficient and accurate recognition of green target fruit, and performs better in terms of generalization ability and robustness under the interference of complex orchard environment. In future research, more complex situations in orchards and the efficiency of the model are further considered on the assembly capability and real-time operational capability of the equipment. Given good recognition of green target fruit of the new model, it can be further extended to other fruit and vegetable production.

This model of ours has a relatively fast recognition speed, high accuracy, and strong generalization ability of the model. Although the method has achieved relatively good results so far, it also needs to consider the accuracy and internal consumption in practical problems, and this model still need to continuously improve the efficiency in the future, and optimize the network structure of the model to improve the operation speed and efficiency under the improvement of accuracy.

Declaration of Competing Interest

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

Acknowledgments

This work is supported by Natural Science Foundation of Shandong Province in China (ZR2020MF076, ZR2019ZD04); Focus on Research and Development Plan in Shandong Province (No.: 2019GNC106115); National Nature Science Foundation of China (No.: 62072289, 81871508); Shandong Province Higher Educational Science and Technology Program (No.: J18KA308); Taishan Scholar Program of Shandong Province of China (No.: TSHW201502038).

References

Zhang, W., Chen, K., Wang, J., Shi, Y., Guo, W., 2021. Easy domain adaptation method for filling the species gap in deep learning-based fruit detection. Hortic. Res. 8 (1). https://doi.org/10.1038/s41438-021-00553-8.

Schima, R., Mollenhauer, H., Grenzdörffer, G., Merbach, I., Lausch, A., Dietrich, P., Bumberger, J., 2016. Imagine all the plants: Evaluation of a light-field camera for on-site crop growth monitoring. Remote Sensing 8 (10), 823. https://doi.org/10.3390/rs8100823.

Zhao, Y., Gong, L., Huang, Y., Liu, C., 2016. A review of key techniques of vision-based control for harvesting robot. Comput. Electron. Agric. 127, 311–323.

Bauer, A., Bostrom, A.G., Ball, J., Applegate, C., Cheng, T., Laycock, S., Rojas, S.M., Kirwan, J., Zhou, J., 2019. Combining computer vision and deep learning to enable ultra-scale aerial phenotyping and precision agriculture: A case study of lettuce production. Hortic. Res. 6 (1). https://doi.org/10.1038/s41438-019-0151-5.

Arefi, A., Motlagh, A.M., Mollazade, K., et al., 2011. Recognition and localization of ripen tomato based on machine vision. Aust. J. Crop Sci. 5 (10), 1144–1149.

Dorj, U.-O., Lee, M., Yun, S.-S., 2017. An yield estimation in citrus orchards via fruit detection and counting using image processing. Comput. Electron. Agric. 140, 103–112.

Tian, Y., Duan, H., Luo, R., Zhang, Y., Jia, W., Lian, J., Zheng, Y., Ruan, C., Li, C., 2019. Fast recognition and location of target fruit based on depth information. IEEE Access 7, 170553–170563.

Bhunia, A.K., Bhattacharyya, A., Banerjee, P., Roy, P.P., Murala, S., 2020. A novel feature descriptor for image retrieval by combining modified color histogram and diagonally symmetric co-occurrence texture pattern. Pattern Anal. Appl. 23 (2), 703–723.

Bhattacharyya, A., Saini, R., Roy, P.P., Dogra, D.P., Kar, S., 2019. Recognizing gender from human facial regions using genetic algorithm. Soft. Comput. 23 (17), 8085–8100.

Bargoti, S., Underwood, J.P., 2017. Image segmentation for fruit detection and yield estimation in apple orchards[J]. J. Field Rob. 34 (6), 1039–1060.

Jia, W., Tian, Y., Luo, R., Zhang, Z., Lian, J., Zheng, Y., 2020. Detection and segmentation of overlapped fruits based on optimized mask R-CNN application in apple harvesting robot. Comput. Electron. Agric. 172, 105380.

He K, Gkioxari G, Dollár P, et al. Mask r-cnn. Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.

He, K., Zhang, X., Ren, S., et al., 2016. Identity mappings in deep residual networks. European Conference on Computer Vision, 630–645.

Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4700-4708.

Chen, S.W., Shivakumar, S.S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C.J., Kumar, V., 2017. Counting apples and oranges with deep learning: A data-driven approach. IEEE Rob. Autom. Lett. 2 (2), 781–788.

Gupta, S., Roy, P.P., Dogra, D.P., Kim, B.-G., 2020. Retrieval of colour and texture images using local directional peak valley binary pattern. Pattern Anal. Appl. 23 (4), 1569–1585.

Ghose, S., Chowdhury, P.N., Roy, P.P., et al., 2021. Modeling Extent-of-Texture Information for Ground Terrain Recognition. In: 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, pp. 4766–4773.

Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9627-9636.

Wang X, Girshick R, Gupta A, et al. Non-local neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794-7803.

Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

Veličković P, Cucurull G, Casanova A, et al. Graph attention networks. arXiv preprint arXiv:1710.10903, 2017.

Lin, T.Y., Maire, M., Belongie, S., et al., 2014. Microsoft coco: Common objects in context. European Conference on Computer Vision, 740–755.

Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.

Huang Z, Huang L, Gong Y, et al. Mask scoring r-cnn. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 6409-6418.

Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3146-3154.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, 1097–1105.

Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision. 2017: 2980-2988.

Yu J, Jiang Y, Wang Z, et al. Unitbox: An advanced object detection network. Proceedings of the 24th ACM International Conference on Multimedia. 2016: 516-520.

de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 134 (1), 19–67.

Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning. PMLR, 2015: 448-456.

Bottou, L., 2010. Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT'2010. Physica-Verlag HD, pp. 177–186.

Ren, S., He, K., Girshick, R., et al., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems 28, 91–99.

Liu, W., Anguelov, D., Erhan, D., et al., 2016. Ssd: Single shot multibox detector. European Conference on Computer Vision, 21–37.

Redmon J, Farhadi A. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

Bolya D, Zhou C, Xiao F, et al. Yolact: Real-time instance segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 9157-9166.

Bolya D, Zhou C, Xiao F, et al. Yolact++: Better real-time instance segmentation. arXiv preprint arXiv:1912.06218, 2019.

Wang, X., Kong, T., Shen, C., et al., 2020. Solo: Segmenting objects by locations. European Conference on Computer Vision, 649–665.

Xie, E., Sun, P., Song, X., et al., 2020. Polarmask: Single shot instance segmentation with polar representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12193–12202.

Fu C Y, Shvets M, Berg A C. RetinaMask: Learning to predict masks improves state-of-the-art single-shot detection for free. arXiv preprint arXiv:1901.03353, 2019.

Ke L, Tai Y W, Tang C K. Deep Occlusion-Aware Instance Segmentation with Overlapping BiLayers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4019-4028.