Instance Segmentation Method based on R2SC-Yolact++

Liqun Ma[1], Chuang Cai[2], Haonan Xie[3], Xuanxuan Fan[4], Zhijian Qu[5], Chongguang Ren[6*]
School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China[1, 2, 3, 4, 5, 6]
Jinan Inspur Data Technology Co., Ltd, Jinan 250000, China[6]

***Abstract*—To address the problems of missed detection, segmentation error and poor target edge segmentation in the instance segmentation model, a R2SC-Yolact++ instance segmentation approach based on the improved Yolact++ is proposed. Firstly, the backbone network adopts Res2Net which introduces spatial attention mechanism (SAM) to improve the problem of segmentation error by better extracting feature information; then, high-quality masks are obtained by fusing the detail information of the shallow feature P2 as the input to the prototype mask branch; finally, the problem of missed detection was solved by introducing Cluster-NMS in order to improve the accuracy of the detection boxes. In order to illustrate the effectiveness of the improved model, experiments were conducted on two publicly available datasets, the COCO and CVPPP datasets. The experimental results show that the accuracy on the COCO dataset is 1.1% higher than the original model. And the accuracy on the CVPPP dataset is 1.7% better than before the improvement, which is better than other mainstream instance segmentation algorithms such as Mask RCNN. Finally, the improved model is applied to the insulator dataset, which can segment the shed of insulator accurately.**

***Keywords*—Instance segmentation; Yolact++; Res2net; Cluster-NMS; insulator dataset**

I. INTRODUCTION

With the rapid development of deep learning, our life is also moving towards automation, such as automatic driving and automatic picking, etc. The first problem that needs to be solved for these tasks is to recognize the category as well as the location of the target, and the current deep learning based methods mainly include object detection and image segmentation. Object detection can only detect the category as well as the location of the target, cannot identify the specific and accurate outline of the target. Deep learning-based image segmentation methods include semantic segmentation and instance segmentation. Semantic segmentation can only classify the targets in an image into different categories, and different instances belonging to the same category cannot be distinguished. Instance segmentation can identify both the class and the exact contour position of different instances. Therefore, it is important to study how to accurately segment different instances using instance segmentation.

Instance segmentation is an important and difficult branch of computer vision, and its task is to identify the target contour location and classify it at pixel level to get the segmentation mask of different instances. The existing instance segmentation algorithms are mainly classified into single-stage and two-stage instance segmentation methods. The two-stage instance segmentation method has the advantage of higher accuracy, but the model is complex and slow. Mask RCNN [1], proposed in 2017, is one of the most commonly used detection-based methods, using the two-step idea of Faster RCNN [2], to which a branch of prediction segmentation masks is added. PANet [3] made improvements to Mask RCNN by introducing bottom-up paths and expanding the feature pyramid network, which improved the accuracy, but also slowed down a lot. The advantage of single-stage segmentation methods is that the model is simple and fast, but the accuracy is relatively low. Typical single-stage methods include the anchor frame-based method YOLACT proposed by Bolya et al. [4], which divides the instance segmentation task into two parallel branches and achieves real-time instance segmentation for the first time, but the accuracy was poor, and it was later improved by proposing YOLACT++ [5], which significantly improved the segmentation accuracy by adding deformable convolution and presetting more anchor frames using mask rescoring. The method CondInst [6] was proposed in 2020 to solve instance segmentation from a new perspective, using dynamic masks and not relying on ROI operations, achieving higher accuracy and being faster. Since the accuracy of the anchor frame-based method depends greatly on the set anchor frame hyper-parameters, a segmentation method without anchor frames is proposed afterwards. SOLO [7] proposed in 2020 uses the location information of instances for instance classification. It is to divide the different instances by assigning the instances to different channels based on the fact that each instance has a different center point and size. The accuracy was further improved with the later improved SOLO V2 [8]. In recent years, with the amazing results achieved by Transformer in natural language processing, it has also been applied to instance segmentation with good results [9, 10].

Although there has been a great development in instance segmentation, there is still a lot of room for development of existing models in terms of accuracy improvement. The problems such as segmentation errors as well as poor target edge segmentation due to insufficient extraction of image information, and missed detection due to inaccuracy of prediction boxes, all contribute to the low accuracy of segmentation. To address the above problems, this paper proposes an instance segmentation method based on the improved Yolact++. The method can better extract features, effectively improve the problems of missed detection and segmentation errors, and enhance the accuracy of instance segmentation. The main contributions of this paper are summarized as follows:

- The R2SC-Yolact++ algorithm is proposed to address the problems of segmentation error and missed detection in Yolact++.
- The model introduces Res2Net for the segmentation error problem and embeds a spatial attention mechanism for better feature extraction, and fuses shallow features with P3 as input to the protonet branch to solve the problem of poor segmentation edges, and finally Cluster-NMS was introduced to solve the problem of missed detection caused by the suppression of too many detection boxes.
- Validating and comparing the segmentation performance of the proposed model on two publicly available datasets and a homemade insulator dataset.
- The paper is organized as follows: Section II presents the related works for segmentation, Section III describes the proposed R2SC-Yolact++ method, Section IV analyzes and discusses the experimental results, and Section V concludes the paper.

II. Related Work

Early traditional image segmentation mainly used the digital image processing technology; the main methods include segmentation methods based on region, edge detection, etc. OSTU finds the threshold value as the segmentation value to distinguish the foreground from the background based on the gray scale distribution of the gray scale map using inter-class variance method. Canny [11] edge detection algorithm performs edge detection by finding the optimal edge pixels and is the most commonly used segmentation method. These methods is generally based on the brightness and color of the pixels in the image, as well as the degree of variation in the pixel values, so they are susceptible to segmentation errors due to uneven illumination, noise, and other factors.

With the development of deep learning, segmentation methods based on deep learning are proposed, which are divided into semantic segmentation and instance segmentation. Semantic segmentation is a pixel-level classification of images to distinguish different classes of targets. The beginning of semantic segmentation was the use of full convolutional networks [12] (FCN) for classification, after which semantic segmentation developed rapidly. U-Net [13] was proposed for medical segmentation in 2015, which uses an encoder-decoder structure to extract features and fuse shallow features with deep semantic information to fully use image contextual information for accurate segmentation of medical images, but has the problem of slow segmentation speed. The DeepLab series [14-17] proposed afterwards have continuously improved segmentation accuracy and efficiency by introducing atrous convolution to reduce model parameters, proposing ASPP to solve the problem of target segmentation at different scales, and introducing decoder structure to optimize the problem of poor segmentation edge accuracy. SegFormer [18] proposed a hierarchical Transformer encoder structure by combining semantic segmentation with Transformer, using overlapping patches to ensure the local continuity of features, while using deep convolution to replace position encoding to convey position information and obtain a high-quality segmentation map. But semantic segmentation has a problem in application, it cannot distinguish different instances belonging to the same class, and many applications need to mark each different instance, so later proposed instance segmentation. However, how to improve the accuracy of instance segmentation and accurately segment the target is an urgent problem.

Existing instance segmentation algorithms are continuously improved from all angles to enhance the accuracy of segmentation. For the problem of segmentation errors and missed segmentation, many studies ResNeXt [19], Res2Net [20], etc. are solved from the perspective of better feature extraction to improve the feature expression ability of the backbone network. The use of attention mechanisms [21-23] has also been proposed to make the network focus on important features and suppress unnecessary features. The accuracy of the detection boxes also affects the accuracy of the segmentation, so many scholars have devoted to obtaining more accurate detection boxes by using more comprehensively computed loss functions DIoU, CIoU [24], etc., or by using more appropriate non-maximum suppression to obtain detection boxes that are more in line with the target. Nowadays the algorithms also commonly have the problem of poor target edge segmentation, many solutions have been proposed for this problem, such as adding an additional loss to the segmentation boundary of the network output, and Xuecheng Li [25] proposed an edge loss function so as to improve the accuracy of the segmentation edge. Improving the segmentation of edges by increasing the resolution of the input image as well as the extracted feature maps is also a relatively straightforward approach, but this adds a significant amount of computation.

The Yolact++ instance segmentation model chosen in this paper, which is simple and realizes real-time segmentation, has been applied in many fields and improvements have been made to address the problems so that they can better segment the objects. Yajun Li in [26] proposed an extended network based on Yolact that can simultaneously detect fruit bundles as well as segment fruit stalks, which can accurately and quickly identify the pose of fruit bunches to support successful picking by picking robots. Based on Yolact++, RTLSeg is proposed to segment the railroad track [27], the feature enhancement module is proposed to improve the feature extraction and characterization ability of the model, PaFPN is used to enhance the interaction of information, as well as the location awareness is added to the Protonet module to obtain a high-quality prototype mask. These improvements effectively enhance the segmentation accuracy of the railroad track and can accurately and efficiently segment the railroad track line components. Zhenni Shang [28] used Yolact, which introduces the SE attention mechanism to enhance feature expression and FRelu activation function, for efficient segmentation of protozoa in microscopic images.

III. Model Overview

A. Principle of Yolact++

ResNet50 with deformable convolution was introduced as the backbone network to extract features from the input image, and total of five different scales of feature maps, C1-C5, were obtained. For efficiency, Yolact++ uses only C3-C5 for feature fusion through FPN to obtain P3-P5, and downsamples from

P5 to obtain P6 and P7. Then after two parallel branches, one branch is the detection branch with inputs P3-P7, anchor frames with aspect ratios [1, 1/2, 2, 1/3, 3] are generated and classified, regressed, mask coefficients predicted, and non-maximum suppression (Fast NMS) is applied to get the final instance prediction boxes. The other branch is the prototype mask branch, which uses P3 as input and obtains k mask maps corresponding to different regions after convolution and upsample. The prototype mask maps and the mask coefficients are linearly combined by matrix multiplication to obtain the instance masks, which are intercepted by the prediction boxes, and the final mask segmentation maps are obtained by threshold processing and binarization. The structure of the Yolact++ model is shown in Fig. 1.

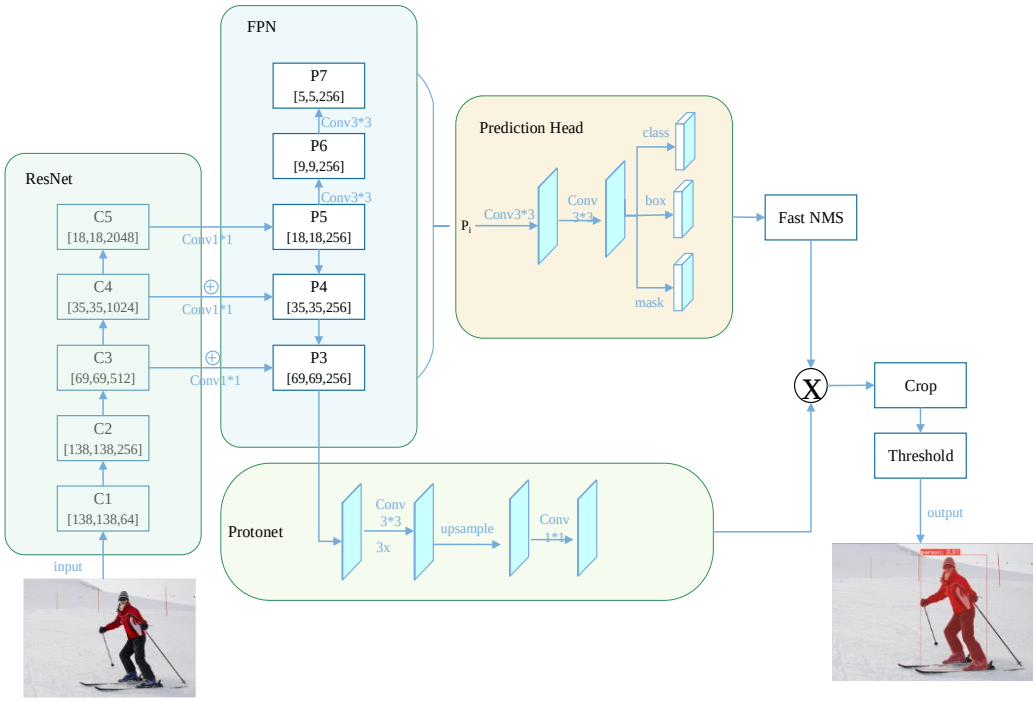

Fig. 1. Structure of the Yolact++ model.

B. Improved Backbone

The quality of feature maps for feature extraction directly affects the subsequent detection and segmentation, so the backbone network is improved to address the problem of inadequate feature extraction. The Res2Net module is used to replace the bottleneck structure of ResNet to represent multi-scale features at a finer granularity, and to increase the receptive field of each network layer to extract more detailed information about the edges as well as the overall information about the target. ResNet bottleneck structure requires layer-by-layer 3×3 convolution for multi-scale feature extraction. Res2Net is realized by a set of (s layered) 3×3 convolutions to do multiple scales within a block, using similar connections to residual networks for connectivity, which can output features with more number of different scales and enhance the expressive power of the network. This is achieved by averaging the feature map obtained after 1×1 convolution into four sub-feature maps with the same number of channels (n/4), the third sub-feature map x_3 is summed with the new sub-feature map y_2 obtained from the second sub-feature map x_2 after 3×3 convolution as an input to the corresponding convolution to obtain the feature map, and so on. In order to fuse the information of different scales, the four feature maps obtained are merged and the final feature map is output after 1×1 convolution. Res2Net is beneficial for extracting global and local information by segmenting and re-splicing the feature maps for combined use. In order to make the network more focused on the target region and reduce the interference from the background, the spatial attention mechanism is added after the 1×1 convolution and before the residual connection, as shown in Fig. 2.

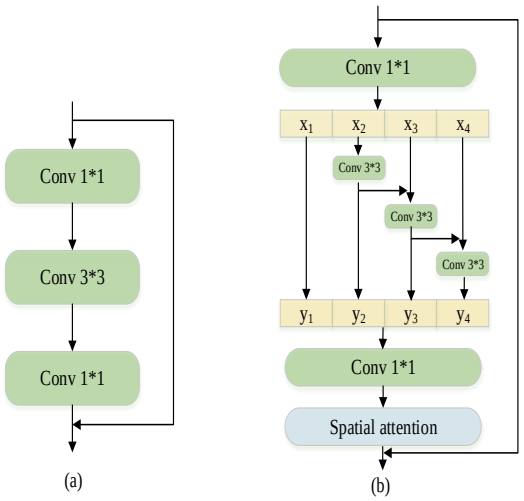

Fig. 2. Comparison of original Resnet bottleneck block and improved res2net module: (a) Resnet bottleneck; (b) Improved res2net module.

Pixels at different locations in an image have different importance, with pixels close to the target being more important and those far from the target more likely to be in the background. The role of spatial attention mechanisms is to enable the network to focus on the "where" part of the information, focusing on important features and suppressing unnecessary features. The specific implementation of the spatial attention mechanism is shown in Fig. 3. Firstly, average pooling and maximum pooling are performed on the feature maps along the channel direction to aggregate the channel information of the feature maps, two H×W×1 feature maps are obtained for stacking, and a feature map is obtained after a 7×7 convolution, and then the spatial attention weight is obtained by the sigmoid function. Finally, the original feature map is used to multiply the spatial weights and output the feature map with positional importance.

C. Prototype Mask Branch Improvement

The higher the resolution of the image, the more detailed information it contains, but the more computationally intensive it is. Yolact++ only uses C3-C5 for feature fusion through FPN in order to raise the speed of computation, and uses the deepest feature map P3 as the input to the Protonet branch. Because of the lack of utilization of shallow features, a lot of detailed information is lost, resulting in poor quality of the resulting prototype masks and poor segmentation of the edges. So in order to get high quality prototype masks, we fuse shallow feature maps as input. Firstly, P2 is obtained by fusing features with FPN, and fusing the low-level information requires convolution of P2 to complete the downsampling. Then P3 is added with the downsampled P2 to get the fused features as the input to the prototype mask branch, and the prototype mask maps are obtained after the protonet branch. The realization process is shown in Fig. 4.

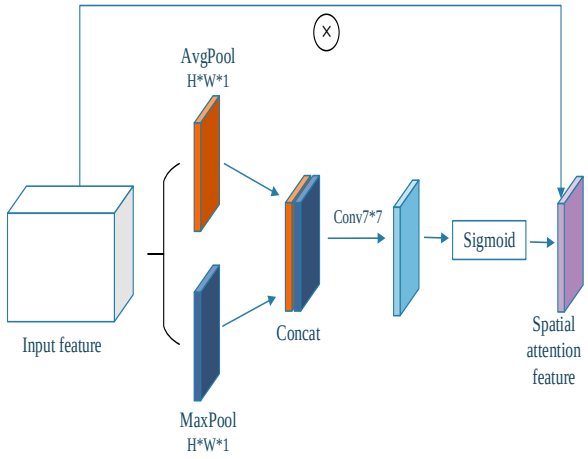

Fig. 3. Spatial attention module.

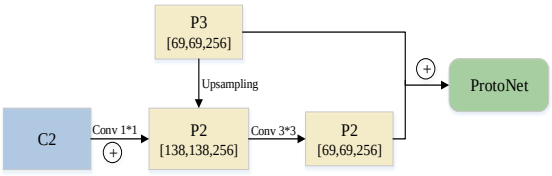

Fig. 4. Structure of the fused shallow feature map.

D. Improvement of Non-Maximum Suppression

The segmentation accuracy of the anchor-based instance segmentation method is affected to some extent by the detection performance. In Yolact++, the segmentation performance is improved by increasing the number of frames to detect difficult and different scale targets, so a large number of redundant boxes are generated, and the model uses Fast NMS instead of Traditional NMS in order to improve the detection speed. It obtains the thresholded binarized one-dimensional vector $b=\{b_i\}_{1\times N}$, $b_i \in \{0, 1\}$ by computing the IoU matrix X for every two boxes, performing upper triangulation, and solving for the maximum value of each column to directly remove the detection boxes with high overlap. However, this leads to the suppression of too many detection frames, and detection boxes with high overlap but belonging to different instances are deleted, leading to the occurrence of missed detection, which reduces the detection accuracy and thus affects the accuracy of segmentation. Therefore, this paper proposes the use of Cluster-NMS instead of Fast NMS, which can make up for the problem of accuracy degradation caused by Fast NMS while ensuring the detection speed. Instead of suppressing the prediction box directly after obtaining the vector b, Cluster-NMS introduces two matrices with the following equations. First, b is expanded into a symmetric matrix with the same diagonal elements as b, and then the matrix C is obtained by left-multiplying the matrix X. The purpose is to omit the effect of the suppressed boxes from the previous iteration on the other boxes. And keep performing the above iterations for C. The end of the iteration is indicated when the b obtained from two consecutive iterations is unchanged. By introducing Cluster-NMS, the problem of missed target detection leading to missed segmentation can be improved and the segmentation accuracy of the model can be improved while maintaining high efficiency.

$$\boldsymbol{A}^{t} = \operatorname{diag}\left(\boldsymbol{b}^{t-1}\right) \tag{1}$$

$$\boldsymbol{C}^{t} = \boldsymbol{A}^{t} \times \boldsymbol{X} \tag{2}$$

IV. Experimental Methods and Analysis of Results

A. Experimental Environment and Parameter Description

All experiments in this paper are based on Windows 10 operating system, hardware system is Intel Core i5-10400F, graphics card is NVIDIA Geforce RTX 2070 SUPER and running memory is 8G. GPU processing six images at a time in ablation experiments, and the size of images are uniformly processed to 550×550. Initial momentum is set to 0.9, learning rate is 0.001, weight decay is 0.0005.

B. Datasets

To validate the effectiveness of the improved model, this paper uses two publicly available datasets for training and testing, and the models were applied to the insulator dataset produced by ourselves to segment the shed of insulator. The MS COCO dataset is a large image dataset developed and maintained by Microsoft and is the most commonly used open standard dataset. In this paper, we conduct comparative experiments of instance segmentation algorithms using the COCO 2017 dataset, which contains 80 categories of life. The

CVPPP dataset is a plant image dataset that provides tobacco and Arabidopsis raw images as well as labeled images to segment plant leaves. The dataset is divided into a total of four sub-datasets, A1-A4, and A5 is the sum of the four sub-datasets, including 810 images of the training set. In order to facilitate subsequent processing and comparison of results, the dataset labels were processed into COCO format. In this paper, ablation experiment are conducted on the CVPPP dataset to verify the effectiveness of the improved module, and the segmentation effect before and after the model improvement is tested on the COCO dataset. And comparison experiments are conducted on CVPPP dataset to compare and analyze the segmentation results of different models.

We need to segment the shed of insulator to find the shed location, so we made insulator dataset and segmented it using improved algorithm. Most of the collected insulator pictures are composite insulators. In addition, in order to ensure the diversity of insulators and improve the generalization ability of the model, insulator images of different types and environments are collected from the Internet. The images were labeled using LabelMe, and a total of 581 images were labeled and randomly augmented by flipping, panning, adding noise, and changing brightness to 2316 images with 19204 instances. The dataset was divided into a training set of 1481, a validation set of 371, and a test set of 464. To facilitate subsequent processing, the dataset is converted to COCO dataset format. According to the configuration of the experimental environment and to ensure the validity of the experimental comparison, the image size was resized to 550×550 in all experiments.

C. Evaluation Metrics

All experiments in this paper were evaluated and analyzed using COCO evaluation metrics, mainly showing the mAP, AP_{50}, AP_S, AP_M, AP_L. Average accuracy AP is the mean value of accuracy at an IoU of 0.5-0.9, an interval of 0.05, and a recall of 0-1 under a category, calculated as shown in Eq. (3), and the area under a two-dimensional curve plotted with recall as the horizontal axis and precision as the vertical axis. MAP is the mean value of AP for all categories, AP_{50} for accuracy at IoU=0.5. S, M, and L are distinguished according to the size of the area of the examples, and the accuracy is obtained separately.

$$AP = \int_0^1 P(\mathrm{r})\mathrm{dr} \tag{3}$$

D. Ablation Experiments

In order to quantitatively analyze the impact of the introduction of Res2Net in Yolact++ combined with attention to the backbone network, feature fusion, and improved non-maximal value suppression on the segmentation performance, this paper uses the above methods in combination with Yolact++ and ablation experiments on the CVPPP dataset. All experiments on the CVPP dataset were iterated 150,000 times. The structure of each model in the ablation experiment and the results are shown in Table I. Res2Net indicates the backbone module of Res2Net combined with the attention mechanism, feature fusion indicates that P2 is fused with P3 as the input of the prototype mask, and "√" indicates the introduction of the corresponding module. The experimental results quantify the effects of these methods on plant leaf segmentation performance.

TABLE I. COMPARISON OF ABLATION EXPERIMENTAL RESULTS OF CVPPP DATASET

Experiment	Res2net	Feature Fusion	Cluster-NMS	mAP
1	-	-	-	65.6
2	√	-	-	66.7
3	√	√	-	66.9
4	√	-	√	67.2
5	√	√	√	67.3

Experiment 1 shows the segmentation results of the baseline model on the CVPPP dataset. In order to harmonize the evaluation criteria, the paper does not use the evaluation metrics specified in the CVPPP dataset, but instead uses the COCO evaluation metrics. Experiment 2 was the improved Res2net combined with Yolact++, and the mAP was improved by 1.1%, comparing with Experiment 1 shows that the introduction of Res2net is more beneficial to extract global as well as local features, and the addition of spatial attention can make the network focus more on the target part, thus improving the segmentation accuracy. Experiment 3 represents the fusion of shallow features with the original Protonet input feature P3 based on Experiment 2 to increase the feature detail information and get a higher quality prototype mask, which results in more accurate edge segmentation. Experiment 4 is the introduction of Cluster-NMS on the basis of Experiment 2. Because Fast NMS is prone to treat two frames that are close to each other but belong to two different similar instances as two overlapping frames of one instance, one of the frames will be suppressed for deletion, and thus there will be a problem of missed detection. And the leaves of a plant are similar in shape and close together, so it is more likely to have the problem of missed detection. Compared with Experiment 2, mAP is improved, indicating that Cluster-NMS can solve the problem of Fast NMS suppressing too many boxes and leading to missed detection, thus improving the segmentation accuracy. Experiment 5 shows the segmentation results of our model, which combines the above method with Yolact++, and the experimental results show that our model improves the segmentation accuracy of plant leaves by 1.7% compared to the accuracy before improvement.

Fig. 5 visualizes the variation of the total loss value over the training set, which serves as a measure of the error between the predicted results and the true annotation, with smaller losses indicating more accurate predictions and better segmentation performance of the model. As shown in the figure, the loss values of the baseline model Yolact++ as well as the improved model are visualized in the figure. The loss values of the two models leveled off after about 110,000 iterations, with the pre-improvement model eventually stabilizing at about 1.41 and the post-improvement model loss values eventually stabilizing at about 1.35. In general, the loss value curves of the improved model are below those of the baseline model. As well as Fig. 6 reflects the mAP comparison on the validation set before and after the model improvement,

both of which demonstrate that the segmentation performance of the improved model is better than that of the baseline model.

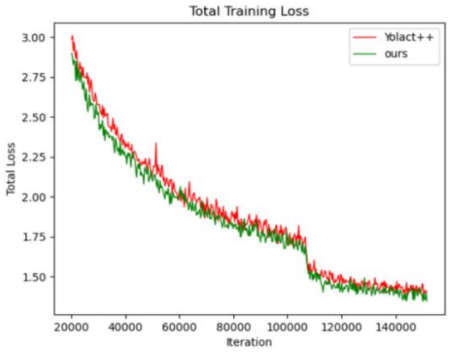

Fig. 5. Loss value curve.

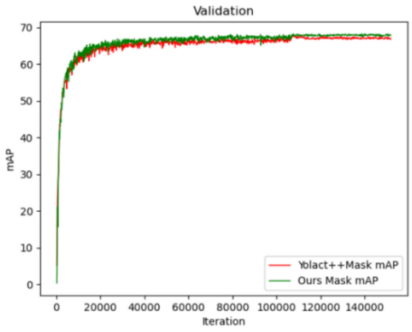

Fig. 6. CVPPP validation set mAP plot.

Fig. 7 visualizes the leaf segmentation results before and after the model improvement. Column (a) of the figure shows the original image of the plant leaf, and the red boxes in the figure show where the results are significantly better after the improvement. Columns (b) and (d) show the leaf segmentation results for the model before and after the improvement, respectively, and column (c) shows the leaf segmentation results obtained from Experiment 2 in the ablation experiment. There are two shaded leaves segmented into one leaf in the red box of the first image, and there are cases where the edge portion is not accurately segmented. After feature extraction by Res2net combined with spatial attention, the features are better extracted and the edges are segmented accurately. In column (d) all leaves are accurately segmented. Comparing columns (b) and (c) of the second figure illustrates that the improved backbone network can solve the problem of segmentation error, but there is still the problem of missed segmentation, comparing column (c) and (d) shows that Cluster-NMS can solve the problem of missed segmentation. The missed detections in the third figure are also well resolved, indicating that our model has better segmentation than Yolact++.

Table II shows the segmentation performance of the model before and after improvement on the MS COCO val2017. The experiment set batch_size to 4 and trained 54 epochs. Because the background of the COCO dataset is more complex and the number and size of targets differ significantly from the CVPPP dataset, resulting in a lower mAP of the mask than the CVPPP dataset. And due to the problem of different computer configurations, the experimental results are somewhat different from those of the published papers. The table shows that the accuracy of the model is 1.1% higher after the model improvement than before the improvement, and the mAP of the instances of different sizes is improved.

TABLE II. SEGMENTATION RESULTS ON COCO DATASET

Model	Backbone	mAP	AP_{50}	AP_S	AP_M	AP_L
Yolact++	Resnet50	33.1	51.8	11.6	35.5	54.3
Ours	Resnet50	34.2	52.8	12.1	36.9	56.1

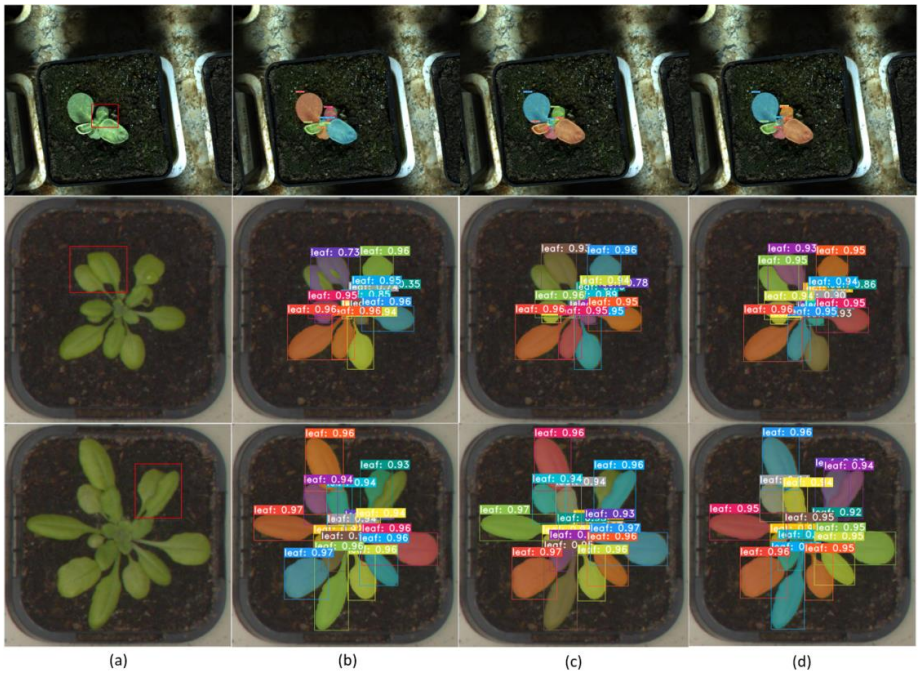

Fig. 7. Visualization of leaf segmentation results.

E. Comparison of Different Models

To further validate the effectiveness of the improved model, we trained the CVPPP dataset and insulator dataset using several commonly used segmentation models, and tested and compared the segmentation results, as shown in Table III and Table IV. The tables show the mAP of the two-stage segmentation model Mask RCNN, the single-stage models CondInst, Yolact++, the segmentation models without anchor boxes SOLO, SOLO V2, and the improved algorithm of this paper on two datasets, as well as the AP_{50} and the accuracy of these models for different sized targets (AP_S, AP_M, AP_L). From the experimental results in Table III, the segmentation results of SOLO are poor because the leaf size of a plant is relatively similar and the leaves have small intervals or even overlap each other, so there will be multiple leaves of similar size appearing in the same grid, resulting in poor segmentation of leaves. Other models perform well on the CVPPP dataset, but all of them also have the problem of segmentation error and missed detection. The improved algorithm makes improvements to address the above problems, and the segmentation of plant leaves is better than other models.

Insulators play an important role in the power system, and damage to insulators can affect the entire line. They are an important concern in power system inspections, so it is important to ensure the cleanliness and integrity of insulators. First of all, it is necessary to identify the shed of insulator and find the exact location of the shed. Due to the uneven illumination and shadows caused by the insulator usage scenario, the shed of insulator cannot be accurately segmented by the traditional image segmentation method. So how to quickly and accurately segment a single shed is a problem that needs to be solved urgently, in this paper, we use several instance segmentation algorithms to train insulator datasets and segment the shed of insulator, the results are shown in Table IV, and the segmentation results were visualized as shown in Fig. 8.

TABLE III. COMPARISON OF SEGMENTATION RESULTS OF MAINSTREAM MODELS FOR CVPPP DATASET

Model	Backbone	mAP	AP_{50}	AP_S	AP_M	AP_L
Mask RCNN	Resnet50	62.7	89.2	44.9	79.9	71.7
CondInst	Resnet50	63.5	92.3	43.3	82.5	86.8
SOLO	Resnet50	59.2	83.0	33.3	79.7	79.3
SOLO V2	Resnet50	62.2	83.3	36.3	82.3	83.7
Yolact++	Resnet50	65.6	88.9	45.1	81.5	73.7
Ours	Resnet50	67.3	91.0	45.9	83.3	79.9

TABLE IV. COMPARISON OF THE RESULTS OF DIFFERENT MODELS FOR INSULATOR DATASET

Model	Backbone	mAP	AP_{50}	AP_S	AP_M	AP_L
Mask RCNN	Resnet50	36.3	83.0	24.3	32.0	41.0
CondInst	Resnet50	26.1	75.3	11.3	19.0	34.3
SOLO	Resnet50	37.2	87.8	18.5	32.2	42.7
SOLO V2	Resnet50	40.4	92.3	19.2	34.7	46.8
Yolact++	Resnet50	40.2	78.5	24.7	33.5	47.6
Ours	Resnet50	41.7	82.3	24.5	34.3	49.9

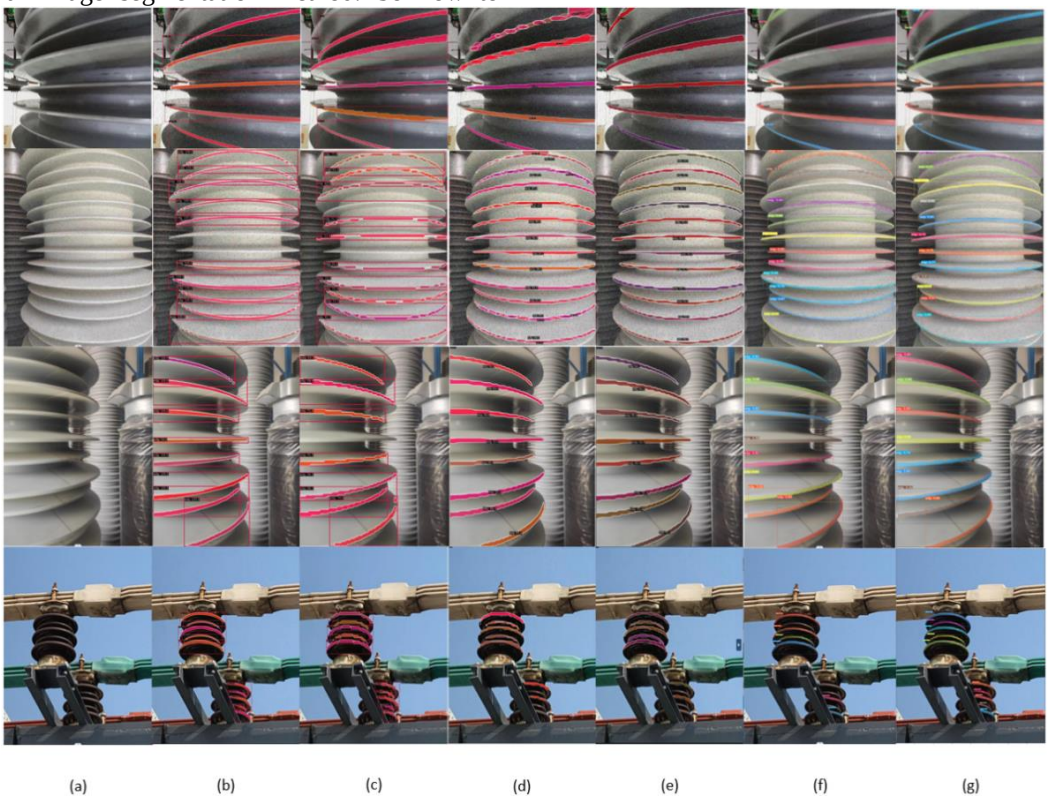

Fig. 8. Comparison of segmentation results of mainstream models: (a) Original image; (b) Mask RCNN segmentation results; (c) CondInst segmentation results; (d) SOLO segmentation results; (e) SOLO V2 segmentation results; (f) Yolact++ segmentation results; (g) Our model segmentation results.

The insulator dataset is used to train the above mainstream example segmentation model and this paper's algorithm, respectively, and the trained model is used to verify the detection and segmentation effect of the shed of insulator. The training results are compared in Table IV. The experimental results show that CondInst has the worst segmentation effect. Mask RCNN, SOLO performs slightly worse on the insulator dataset, mAP is slightly lower than the baseline model, and SOLO V2 and Yolact++ segmented well. The masks of Yolact++ use all the information of the picture space without repooling, which does not cause the performance loss of the mask, the segmentation accuracy for large target objects is significantly higher than other methods, and most of the insulator datasets are large targets, so the baseline model Yolact++ has a higher mAP than other models. The improved model made improvements to the problems of the baseline model, and the accuracy was improved by 1.5% over the baseline model, with the best segmentation results.

Fig. 8 visualizes the shed of insulator segmentation results of the mainstream instance segmentation models. In general, all the models in Table IV have the problems of missed detection and poor segmentation of shed edges. Mask RCNN has obvious missed detection. The most problematic segmentation result of SOLO is the segmentation error, which divides one shed into multiple sheds. SOLO V2 performs better than SOLO, and the problem of segmentation error is well solved, but there is also the problem of poor segmentation edge. The segmentation results of our improved model are the best, and the problems of missed detection as well as poor segmentation edges that occur in the baseline model are improved to segment the shed of insulator more accurately.

F. Discussion

In this paper, segmentation performance is validated and compared on two publicly available datasets as well as a homemade insulator dataset. In order to more easily compare and illustrate the segmentation effects of the analytical models on different datasets, the article uniformly uses the COCO evaluation metrics. Experimental results show that our model achieves better segmentation results on all three datasets. The improvement is greater on the CVPPP and insulator datasets, which have only one category, similar and densely distributed targets, and the algorithm's backbone and NMS method improvements both lead to improved segmentation performance for these targets. The experimental results of the three datasets can also illustrate that the model has a greater improvement in the segmentation of large targets. However, the model focuses on improving accuracy and does not focus on model complexity as well as speed issues, and subsequent studies will further optimize the model. And the collected insulator images have the problem of a single type and scene, and the insulator segmentation dataset needs to be enriched subsequently.

V. CONCLUSION

In this paper, Yolact++ is used as a baseline model to improve the accuracy of the model by aiming at the problems of segmentation error, missed segmentation and low accuracy of edge segmentation. Firstly, the use of Res2net which incorporates a spatial attention mechanism as a backbone enables the backbone network to better extract global and local features, acquire target information with clear boundaries, and improve the problem of segmentation errors. Second, the shallow features P2 and P3 are fused and input to the Protonet branch to obtain a high-quality prototype mask map that does not depend on a particular instance. Finally, the introduction of Cluster-NMS improves the problem of missed detection due to the suppression of too many detected boxes by iteratively and gradually eliminating the influence of the boxes with too much overlap on the other boxes, which improves the detection accuracy and segmentation accuracy. Training and testing on the public datasets COCO and CVPPP datasets are performed to verify the effectiveness of the improved model, and the average accuracy of the mask is 1.1% higher on the COCO dataset, and the accuracy on the CVPPP dataset is 1.7% higher than before the improvement. The improved model is applied to the insulator dataset labeled by itself to segment the shed of insulator, and the experimental results show that a more accurate the shed of insulator segmentation is achieved.

ACKNOWLEDGMENT

Thanks are given for the support of Youth Innovation Team Development Plan of Shandong Province Higher Education (2019KJN048).

REFERENCES

[1] He K M, Gkioxari G, DOLLÁR P, Girshick R, “Mask R-CNN,” Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Oct 22-29, 2017, Washington: IEEE Computer Society, pp. 2980-2988, 2017.

[2] Ren S Q, HE K M, Girshick R B, Sun J, “Faster R-CNN:towards real-time object detection with region proposal networks,” Proceedings of the Annual Conference on Neural Information Processing Systems 2015, Montreal, Dec 7-12,2015, Red Hook: Curran Associates, pp. 91-99, 2015.

[3] LIU S, QI L, QIN H, et al., “Path aggregation network for instance segmentation,” Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-22, 2018, Washington: IEEE Computer Society, pp. 8759-8768, 2018.

[4] Bolya D, Zhou C, Xiao F, Lee Y J, “YOLACT:real-time instance segmentation,” Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Piscataway: IEEE, pp. 9156-9165, 2019.

[5] Bolya D, Zhou C, Xiao F, Lee Y J, “YOLACT++:better real-time instance segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 44(2):1108-1121.

[6] Tian Z, Shen C, Chen H, “Conditional convolutions for instance segmentation,” Computer Vision – ECCV 2020. Berlin, German:Springer, pp. 282-298, 2020.

[7] Wang X, Kong T, Shen C, Jiang Y, Li L, “SOLO:segmenting objects by locations,” LNCS 12363:Proceedings of the 16th European Conference on Computer Vision, Cham:Springer, pp. 649-665, 2020.

[8] Wang X, Zhang R, Kong T, Li L, Shen C, “SOLOv2: Dynamic and fast instance segmentation,” Advances in Neural Information Processing Systems, 2020, 33: 17721-17732.

[9] Guo R, Niu D, Qu L, Li Z, “SOTR:Segmenting Objects with Transformers,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, USA:IEEE, pp. 7157-7166, 2021.

[10] Li F, Zhang H, Xu H, et al., “Mask DINO: Towards A Unified Transformer-based Framework for Object Detection and Segmentation.” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023:3041-3050.

[11] Canny, J., A Computational Approach To Edge Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 8: 679-714, 1986.

[12] Long J, Shelhamer E, Darrell T, "Fully convolutional networks for semantic segmentation," proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

[13] Ronneberger O, Fischer P, Brox T, "U-net:Convolutional networks for biomedical image segmentation," proceedings of the International Conference on Medical image computing and computer-assisted intervention, 2015. Springer.

[14] Chen L-C, Papandreou G, Kokkinos I, K Murphy , AL Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," Computer Science, 2014(4):357-361.

[15] Chen L-C, Papandreou G, Kokkinos I, K Murphy , AL Yuille, "Deeplab:Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4):834-48.

[16] Chen L-C, Papandreou G, Schroff F, Adam H, "Rethinking atrous convolution for semantic image segmentation," Computer Vision and Pattern Recognition, 2017, 17(6): 1-14.

[17] Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H, "Encoder-decoder with atrous separable convolution for semantic image segmentation," proceedings of the Proceedings of the European conference on computer vision(ECCV), 2018.

[18] Xie E, Wang W, Yu Z, et al., "SegFormer:Simple and efficient design for semantic segmentation with transformers," Advances in Neural Information Processing Systems, 2021, 34:12077-90.

[19] Xie S, R Girshick, P Dollár, Tu Z, He K, "Aggregated Residual Transformations for Deep Neural Networks," 2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 5987-5995, 2017.

[20] Gao S, Cheng M, Zhao K, et al., "Res2net:A new multi-scale backbone architecture," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 43(2):652-662.

[21] Hu J, Shen L, Sun G, "Squeeze-and-excitation networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141, 2018.

[22] Woo S, Park J, Lee J Y, Kweon S, "Cbam: Convolutional block attention module," Proceedings of the European con-ference on computer vision (ECCV), pp. 3-19, 2018.

[23] Zhang H, Zu K, Lu J, Zou Y, Meng D, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," Proceedings of the Asian Conference on Computer Vision, Macau, Dec 4, pp. 1161-1177, 2022.

[24] Zheng Z, Wang P, Liu W, et al., "Distance-IoU loss:Faster and better learning for bounding box regression," US:AAAI, 2020, 34(7):12993-13000.

[25] Li X, Zheng Y, Zang M, Jiao W, "Wavelet transform and edge loss-based three-stage segmentation model for retinal vessel," Biomedical Signal Processing and Control, vol. 86, 105355, 2023.

[26] Li Y, Feng Q, Liu C, et al., "MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting," European Journal of Agronomy, vol. 146, 126812, 2023.

[27] Wei D, Wei X, Tang Q, et al., "RTLSeg: A novel multi-component inspection network for railway track line based on instance segmentation," Engineering Applications of Artificial Intelligence, vol. 119, 105822, 2023.

[28] Shang Z, Wang X, Jiang Y, Li Z, Ning J, "Identifying rumen protozoa in microscopic images of ruminant with improved YOLACT instance segmentation," Biosystems Engineering, vol. 215, pp. 156-169, 2022.