Instance Segmentation Method Based on DPA-SOLOV2

Yuyue Feng, Liqun Ma,Yinbao Xie, Zhijian Qu
School of Computer Science and Technology, Shandong University of Technology, Zibo 255049, China

***Abstract*—To solve the problems of missed detection, segmentation errors in instance segmentation models, we propose an instance segmentation approach, DPA-SOLOV2, based on the improved segmenting objects by locations V2 (SOLO V2). Firstly, DPA-SOLOV2 introduces deformable convolutional networks (DCN) into the feature extraction network ResNet50. By freely sampling points to convolve features of any shape, the network can extract feature information more effectively. Secondly, DPA-SOLOV2 uses the path aggregation feature pyramid network (PAFPN) feature fusion method to replace the feature pyramid. By adding a bottom-up path, it can better transmit the location information of features and also enhance the information interaction between features. To prove the effectiveness of the improved model, we conduct experiments on two public datasets, COCO and CVPPP. The experimental results show that the accuracy of the improved model on the COCO dataset is 1.3% higher than that of the original model, and the accuracy on the CVPPP dataset is 1.5% higher than that before the improvement. Finally, the improved model is applied to the insulator dataset, which can accurately segment the umbrella skirt of insulators and outperforms other mainstream instance segmentation algorithms such as Yolact++.**

Keywords*—Instance segmentation; segmenting objects by locations V2; deformable convolutional networks; path aggregation feature pyramid network; insulator dataset

I. INTRODUCTION

Deep learning methods possess excellent performance in the field of object detection and have been widely applied in fields such as autonomous driving, intelligent transportation, national defense security [1-3]. Driven by massive amounts of data, deep learning-based object detection methods can learn features with stronger semantic representation capabilities through the feature extraction network. At the same time, during the forward propagation process of the neural network, redundant calculations of a large number of windows are avoided. While the overall detection speed is improved, the detection accuracy is also significantly enhanced. However, although object detection can locate and classify targets, it is difficult to obtain the precise contours of the targets. Image segmentation based on deep learning includes semantic segmentation and instance segmentation. Semantic segmentation can only divide the targets in an image into different categories, but it cannot distinguish different instances of the same category.

Instance segmentation is an important and challenging task in computer vision. It not only needs to identify the target location but also classify it at the pixel level to obtain the segmentation masks of different instances, thus accurately identifying the category and contour information of different instances. Instance segmentation algorithms are mainly divided into two-stage and one-stage methods. Among them, Mask R-CNN [4] and its improved networks adopt the two-stage method of Faster R-CNN [5]. They detect the target area through candidate boxes, then fine-tune these candidate boxes, and finally perform classification in each candidate box to generate bounding boxes and target masks. The two-stage method can improve the segmentation accuracy, but it relies on multiple branches and a large amount of parameter calculation, making real-time segmentation difficult.

One-stage instance segmentation methods feature simple model structures and fast inference speeds. Currently, the mainstream instance segmentation methods are divided into two categories: anchor-based methods and anchor-free methods. Anchor-based instance segmentation methods group pixels into a set of candidate masks in the image, and then generate the final instance masks through embedding, aggregation, and combination. Bolya proposed an anchor-based method, Yolact[6] that can divide the instance segmentation task into two parallel branches and achieves real-time instance segmentation for the first time but its accuracy is relatively poor. Subsequently, Yolact++ [7] was proposed to address the above issues. By adding deformable convolutions, presetting more anchor boxes, and using mask re-scoring, the segmentation accuracy has been significantly improved. Later, CondInst [8] uses dynamic masks and does not rely on ROI operations, achieving higher accuracy and faster speed. The segmentation accuracy of the aforementioned anchor-based instance segmentation methods depends significantly on the precision of detection boxes, which in turn relies heavily on parameters such as the scale and size of pre-set anchor boxes. Many studies aim to improve detection accuracy by increasing the number of anchor boxes, which not only elevates computational overhead but also tends to cause an imbalance between positive and negative samples. Therefore, anchor-free methods were proposed later. SOLO [9] utilizes the location information of instances for instance classification. Since each instance has a different center point and size, SOLO distinguishes different instances by assigning each instance to a different channel. Subsequently SOLO V2 [10] improves accuracy and speed through decoupling design and Matrix NMS, but still requires substantial computational resources, and there is still room for optimization in target detection performance. In recent years, with the excellent achievements of the Transformer in natural language processing, it has also been applied to instance segmentation and achieved good results [11, 12].

Although instance segmentation technology has made great progress, there is still much room for improvement in the segmentation accuracy of existing models. Issues such as

segmentation errors and missed target detections caused by insufficient extraction of image feature information all lead to a relatively low segmentation accuracy of the models. To address the above problems, this study proposes an instance segmentation method based on the improved SOLO V2. This method can extract image features more comprehensively, effectively alleviate the problems of segmentation errors and missed target detections, and improve the accuracy of instance segmentation. The main contributions of this study are summarized as follows:

- The DPA-SOLOV2 algorithm is proposed to solve the problems of segmentation errors and missed target detections in SOLO V2.
- Deformable convolution is introduced into the model. By convolving features of any shape with free sampling points, the network can extract feature information better. Moreover, the PAFPN feature fusion method is used to replace the feature pyramid. By adding a bottom-up path, the position information of features can be transmitted better, and the information interaction between features is enhanced.
- The segmentation performance of the proposed model is verified and compared on two publicly available datasets and a self-made insulator dataset.

The rest of this study is organized as follows: Related work on instance segmentation and existing problems are described in Section II. Next, we introduce the details of our solution in Section III. Subsequently, we design experiments on the COCO dataset, CVPPP dataset, and insulator dataset to evaluate our model, and present the experimental results and analysis in Section IV. Finally, we conclude our study in Section V.

II. Related Work

The instance segmentation method based on deep learning solves the problem in semantic segmentation that different instances within the same category cannot be distinguished. FCIS [13], BlendMask [14], Mask R-CNN etc. adopts a top-down two-stage segmentation method and Mask R-CNN determines the relationship between pixels and objects within a proposed region. It uses Fast R-CNN for object detection and performs the instance segmentation task by adding a segmentation branch. Based on Mask R-CNN, the literature [15] employs a lightweight backbone network to reduce the number of network parameters and compress the model size. By optimizing the convolutional structure of the Feature Pyramid Network (FPN) and the backbone network, the feature information between the high-level and low-level structures can be completely transmitted. The literature [16] introduces a bottom-up path and an attention mechanism based on Mask R-CNN for object detection and segmentation. Two-stage instance segmentation methods have relatively excellent segmentation accuracy. However, the segmentation speed makes it difficult to meet the requirements of the current application scenarios.

In recent years, to reduce the complexity of instance segmentation methods and improve the target segmentation performance without increasing the complex computational load, Bolya proposed a bottom-up and one-stage segmentation method Yolact. It is improved based on RetinaNet. The prototype mask of each image is generated through the proton network, and at the same time, k mask coefficients are obtained by predicting each target instance and the bounding box. The prediction results of the category branch and the mask branch need to be superimposed according to the coefficients, which has the problem of relatively low accuracy. On this basis, to improve the segmentation accuracy, Shang [17] used Yolact. It introduced the SE attention mechanism to enhance the feature expression and used the FRelu activation function for the efficient segmentation of protozoa in microscopic images. Li proposed an extended network based on Yolact [18], which can detect fruit clusters and segment fruit stalks simultaneously to support the successful picking of the picking robot. The above-mentioned one-stage segmentation methods require the setting of anchor boxes. The segmentation accuracy of anchor-based methods largely depends on the hyperparameters of the set anchor boxes. Many studies generally increase the number of anchor boxes to achieve more accurate detection. However, doing so will increase a large amount of computational load.

To address the above issues, Wang proposed the anchor-free instance segmentation framework SOLO, which realizes instance segmentation by leveraging the idea of semantic segmentation and transforms the instance segmentation problem into two concurrent problems of category prediction and instance mask prediction. SOLO divides an image into S×S grids. It assigns instances to different channels based on the fact that each instance has a distinct center point and size, enabling the differentiation of various instances. However, if the targets in the image are too densely packed, there may be multiple instances appearing in the same grid, which will lead to poor segmentation performance.

The SOLO V2 model was proposed to address the issues existing in the SOLO model. Firstly, the mask prediction is decoupled into the prediction of the convolutional kernel and the learning of the feature map. Additionally, Matrix NMS is proposed, which enables the process that must be traditionally and sequentially implemented in non-maximum suppression to be completed at once through parallel operations, thus improving the efficiency of the model. This model is simple and can achieve real-time segmentation. Therefore, this study selects the SOLO V2 model as the baseline model. Moreover, the model features a simple architecture and can achieve real-time segmentation, making it widely applied by numerous scholars in various fields. Based on SOLO V2, Liu improved the segmentation efficiency of tomato leaf disease areas by improving the feature extraction network and introducing deformable convolution and other methods [19]. MSIS [20] conducts multispectral instance segmentation based on SOLO V2. It has improved the instance segmentation performance of electrical equipment by introducing methods such as the feature fusion module. FPN-DenseNet-SOLO [21] takes the SOLO V2 as the backbone framework, uses the optimized DenseNet-169 as the backbone network, and combines it with the feature pyramid network. It detects and segments instances on the semantic branch and the mask branch, achieving accurate segmentation of poultry under normal and heat stress conditions. This study presents an enhanced model of SOLOV2 named DPA-SOLOV2. By integrating DCN and PAFPN, the model

strengthens the capability of feature information extraction, thus effectively mitigating the common issues of target missed detection and segmentation errors in instance segmentation tasks.

III. Model Overview

A. Principle of SOLO V2

SOLO V2 uses ResNet50 as the backbone to extract features and obtains five feature maps. It takes Stage2 to 5 as the input of the feature pyramid network (FPN) for feature fusion. Deformable convolution network (DCN) is applied in Stage3 to 5, while the original convolutional operations are retained for the rest parts. Finally, a total of five feature maps, namely P2 to P6, are obtained for subsequent operations in several branches. The category branch is responsible for predicting the probability that an instance falling within this grid belongs to each category. The output dimension is S×S×C, where S×S represents the maximum number of instances and C is the number of categories. The mask branch is divided into two parallel branches: the mask kernel branch and the mask feature branch. The convolution kernels and feature maps are generated dynamically. The mask kernel branch is responsible for generating the convolution kernel G according to the number of instances, with an output dimension of S×S×D, where D is the weight of the convolution kernel corresponding to its size. P2 to P5 are used as the inputs of the feature branch. They are resized to the same size and then added together to generate the feature map F. The instance map of the mask branch is generated through dynamic convolution between the generated convolution kernel G and the feature map F. Finally, the final instance mask map is selected through matrix non-maximum suppression (Matrix NMS). The structure of the SOLO V2 model is shown in Fig. 1.

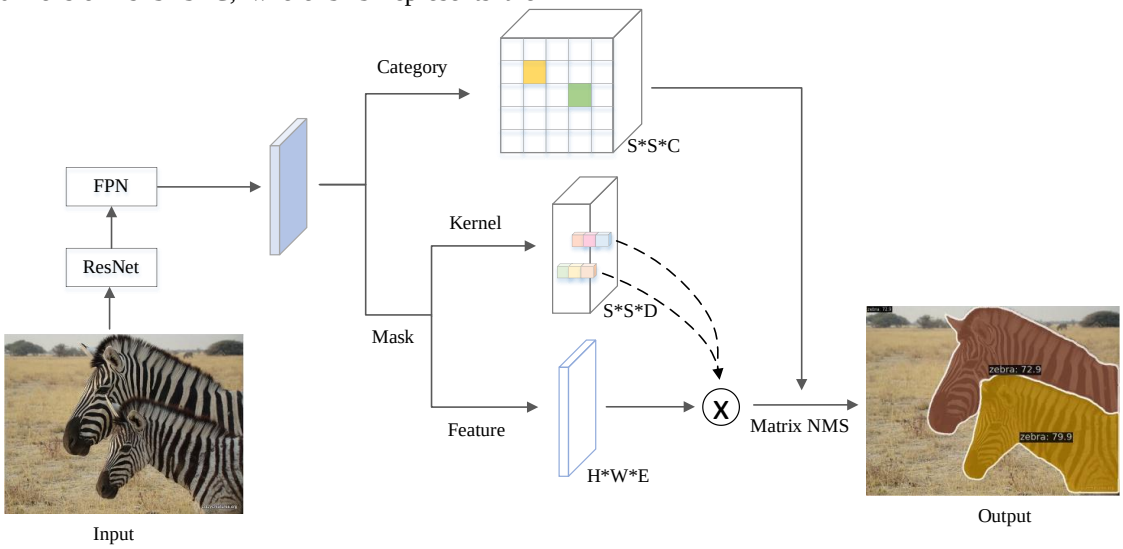

Fig. 1. Structure of the SOLO V2 model.

B. Introduction of Deformable Convolution Networks

The quality of the feature maps used for feature extraction directly affects subsequent detection and segmentation. Therefore, deformable convolution is introduced into ResNet50 to address the problem of insufficient feature extraction. By convolving features of any shape with freely sampled points, the network is enabled to extract feature information better. Ordinary convolution can only extract features using a fixed kernel size and shape. However, most targets are irregular and vary in size, so ordinary convolution has certain limitations in extracting features from irregularly shaped targets. The DCN was proposed mainly to address the issue that the convolution ability of ordinary convolution is affected by spatial transformation. Instead of changing the shape of the convolution kernel, deformable convolution changes the shape of the sampling points of the convolution by adding a position offset to each sampling point.

Fig. 2 shows the difference between ordinary convolution and deformable convolution. From the comparison in the figure, it can be seen that the convolution operation changes from a (ordinary convolution) to irregular sampling point patterns (b and c). DCN can learn any spatial shape of the target through flexible sampling points.

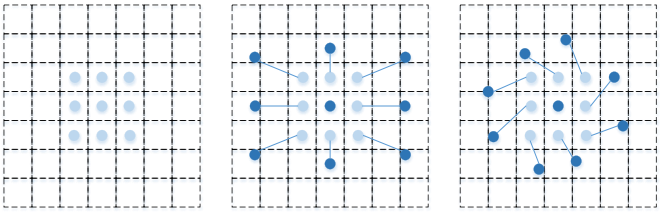

Fig. 2. Normal convolution vs. deformable convolution.

The formula for performing deformable convolution on the sampling point p_0 in the input feature map x is shown in Eq. (1). Here, y represents the output feature map, p_0 is the center point of the convolution kernel, R defines the size and stride of the convolution kernel, w is the weight, p_n is the position of other points in the convolution kernel relative to the center point, Δp_n is the offset of the sampling point, and $x(p_0 + p_n + \Delta p_n)$ calculates the coordinates of each pixel iteratively.

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \quad (1)$$

The convolution operation of DCN is shown in Fig. 3. First, the extracted feature map is taken as the input and passed through convolution for learning, obtaining 2N (where N is the size of the convolution kernel, and each block of the convolution kernel has offset coordinates x and y) offsets for the deformable convolution. The network can learn the weights and offsets of the convolution kernel simultaneously, directly combining the position offsets with the features. Since the offsets are not necessarily integers, bilinear interpolation is used to address this issue. The output feature map is then obtained and used as the input for the next layer.

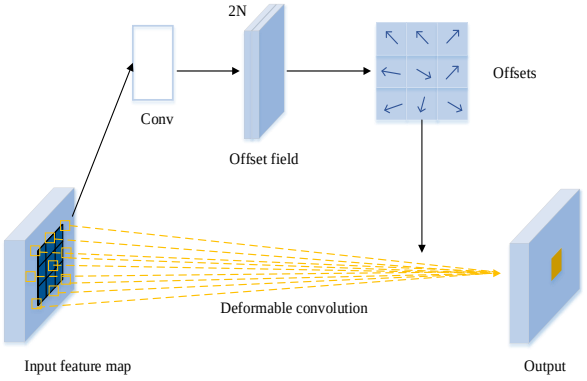

Fig. 3. Diagram of the deformable convolution process.

DCN can address the issue that ordinary convolution is sensitive to spatial transformations such as image translation and rotation. By adaptively adjusting the shape of the convolution kernel, it enhances the invariance to spatial transformations. Moreover, due to its ability to adaptively adjust the shape of the convolution kernel, deformable convolution can better extract feature information of different scales and shapes, thereby improving the performance of the model.

C. Improvement of Feature Pyramid Network

In the forward propagation of neural networks, convolutional operations in multiple layers are required. During this process, the detailed information in the shallow layers is continuously lost, which is not conducive to the propagation of shallow-layer feature information. The feature pyramid network (FPN) can enhance the feature representation of shallow features by transferring semantic information from high-level features to low-level features. The path aggregation feature pyramid network (PAFPN) supplements the FPN by adding a bottom-up path, which enhances the spatial information transfer between features and enables the network to accurately determine the location information of objects. PAFPN introduces a path aggregation module, which allows the network to better integrate multiple feature paths from both bottom-up and top-down directions. The added path only requires a few convolutional layers, enabling the shallow-layer information to be transmitted to the high-layer more quickly and reducing the loss of feature information, thus improving the segmentation accuracy. The structure of PAFPN is shown in Fig. 4. The upper part represents the FPN structure, and a bottom-up path is added on the side. By performing convolution on N_i, the spatial size is reduced to obtain a feature map of the same size as P_{i+1}. Then, the feature map is added pixel-by-pixel to P_{i+1}, to get a new feature map, which endows the feature map with richer feature information. Subsequently, the new feature map is used for classification and mask prediction, which can improve the segmentation accuracy.

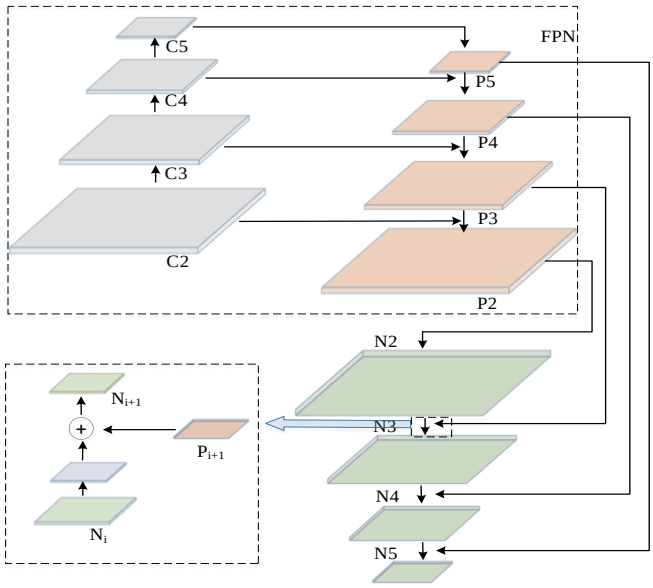

Fig. 4. Structure of PAFPN.

D. Improvement of Non-Maximum Suppression

Matrix NMS is designed for mask suppression. Computing the mask IoU is far more complex than calculating the box IoU. If traditional NMS is used, it will consume a significant amount of time. Therefore, Matrix NMS can substantially save time and improve segmentation efficiency by computing mask IoU in parallel. Inspired by Soft-NMS, Matrix NMS aims to perform parallel operations. Based on this idea, it can complete the suppression process in just one iteration, which greatly reduces the time consumption. Traditional NMS deletes the bounding boxes with lower scores according to the size of the overlapping area. The calculation method is shown in Eq. (2). As a result, the detection and segmentation results are highly susceptible to the set threshold. If the threshold is set too low, the bounding boxes of two adjacent targets may be deleted because the confidence of one box is too small. On the contrary, if the threshold is set too high, the suppression effect will be weak, and false detections are likely to occur. To address this issue, Soft-NMS adopts a smoother approach. Instead of directly deleting the boxes that exceed the set threshold, it first calculates a relatively gentle value to reduce the confidence of these detection boxes. Then, it sorts the remaining boxes according to their scores and finally deletes the boxes with scores lower than the threshold. The filtering results are output after no more boxes are deleted. The formula is shown in Eq. (3), where M represents the box with the highest score and b_i is the adjacent detection box. By multiplying the scores of the detection boxes with excessive overlap by a weight function, the scores of the detection boxes can be attenuated. The larger the IoU of the two boxes is, the

more the score s_i of b_i will decrease. Finally, the detection boxes with scores greater than the set threshold are retained. Since Eq. (3) is a non-differentiable and discontinuous function, it is modified to Eq. (4), where the penalty is greater for the boxes closer to the center of the Gaussian distribution. This approach can retain more boxes and thus improve the accuracy. However, this process can only be carried out sequentially, which requires a large amount of time.

$$s_i = \begin{cases} s_i, iou(M, b_i) < N_t \\ 0, iou(M, b_i) \geq N_t \end{cases} \tag{2}$$

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ s_i(1 - iou(M, b_i)), & iou(M, b_i) \geq N_t \end{cases} \tag{3}$$

$$s_i = s_i e^{-\frac{iou(M, b_i)^2}{\sigma}}, \forall b_i \notin D \tag{4}$$

Soft-NMS can only operate serially, starting from the box with the highest score and iterating step by step. Matrix NMS, on the other hand, focuses on how to parallelize this process. It approaches the problem from the perspective of how a predicted mask m_j is suppressed and proposes using a decay factor to reduce the confidence of the mask. The decay factor is influenced by two aspects. One is the penalty exerted on m_j by all m_i whose scores are higher than that of m_j. The other is the probability that m_i is suppressed. First, the penalty of m_i on m_j needs to be calculated through Eq. (5). However, calculating the probability that m_i is suppressed is not straightforward. Since the probability of a mask being suppressed is generally positively correlated with the IoU, the maximum overlap prediction is directly adopted for approximate calculation, as shown in Eq. (6). Eventually, the calculation process of the decay factor is as presented in Eq. (7), and the updated score is $s_j = s_j \times \text{decay}_j$. The calculation process of Matrix NMS can be completed in one parallel operation. This allows for an improvement in both segmentation accuracy and efficiency. For example, in a complex scene with hundreds of objects, Soft-NMS would take a relatively long time to process each mask sequentially. In contrast, Matrix NMS can handle all these masks simultaneously, reducing the processing time significantly while maintaining or even enhancing the accuracy of the segmentation results.

$$f(iou_{i,j}) = 1 - iou(i, j) \tag{5}$$

$$f(iou_{i,j}) = \min_{\forall s_k > s_i} f(iou_{k,j}) \tag{6}$$

$$decay_j = \min_{\forall s_i > s_j} \frac{f(iou_{i,j})}{f(iou_{.,j})} \tag{7}$$

IV. Experimental Methods and Analysis of Results

A. Experimental Environment and Parameter Description

All experiments in this study are based on the Windows system. MMDetection was used for code construction. PyTorch was selected as the underlying framework to build the model. The GPU was utilized to accelerate the computation by configuring the Cuda and Cudnn environments. The detailed configuration is shown in the following table. 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, and weight decay is 0.0005. The detailed experimental configuration is shown in Table I.

TABLE I. Experimental Configuration Table

Item	Content
CPU	13th Gen Intel(R) Core(TM) i7-13700KF
GPU	NVIDIA GeForce RTX 4090
Video Memory	24GB
Random Access Memory	32GB
Framework	Pytorch1.8.0+cu111
Python	Python3.8

B. Datasets

To verify the effectiveness of the improved model, this study uses two publicly available datasets and an insulator dataset for training and testing. The MS COCO dataset is a large-scale image dataset developed and maintained by Microsoft, and it is the most commonly used open-standard dataset. In this study, we conduct comparative experiments on instance segmentation algorithms using the COCO 2017 dataset, which contains eighty categories of daily items. The CVPPP dataset is a plant image dataset that provides raw images of tobacco and *Arabidopsis thaliana*, as well as labeled images for segmenting plant leaves. This dataset is divided into four sub-datasets, A1-A4 in total, and A5 is the combination of these four sub-datasets, including 810 training set images. To accurately locate the position of the insulator shed, it is necessary to perform segmentation processing on it. Therefore, we have constructed an insulator dataset and used an improved algorithm to segment it. The collected insulator images are mainly composite insulators. Meanwhile, to enhance the diversity of the insulator data and improve the generalization ability of the model, we have also collected insulator images of different types and in various environments from the Internet. These images were labeled using the LabelMe tool, with a total of 581 images being labeled. Subsequently, through data random augmentation methods such as image flipping, translation, noise addition, and brightness adjustment, the number of images was expanded to 2316, containing 19,204 instances. This dataset was further divided into 1,481 training set images, 371 validation set images, and 464 test set images. According to the configuration requirements of the experimental environment, and to ensure the effectiveness of the experimental comparison results, the image size in all experiments was uniformly adjusted to 550×550. To facilitate subsequent processing and result comparison, both the CVPPP dataset and the insulator dataset have been converted into the COCO dataset format.

In this study, ablation experiments are carried out on the insulator dataset to verify the effectiveness of the improved module, and the segmentation effects before and after the model improvement are tested on both the COCO dataset and the CVPPP dataset. Comparative experiments are conducted on the insulator dataset to compare and analyze the segmentation results of different models.

C. Evaluation Metrics

All experiments in this study were evaluated and analyzed using COCO evaluation metrics, mainly showing the mAP, AP_{50}, AP_S, AP_M, AP_L. 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. (8), 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.

$$AP = \int_0^1 P(r)dr \tag{8}$$

D. Ablation Experiments

To quantitatively analyze the impact of introducing the deformable convolution DCN structure and PAFPN on the segmentation ability of the model in SOLO V2, this study combines the above methods with SOLO V2 and conducts ablation experiments on the insulator shed segmentation dataset, specifically the following five ablation experiments:

1) SOLO V2: Conduct the segmentation of insulator shed skirts on the model without any improvements.

2) SOLOV2-DCN: Introduce the deformable convolution structure into the feature extraction network ResNet50 used in the model.

3) SOLOV2-PAFPN: Replace the originally used FPN with the PAFPN structure for feature fusion.

4) The final improved model that integrates the above methods.

5) All of the above methods adopt transfer learning with the pre-trained weights that are trained on the COCO dataset for 1x (12 epochs). In the fifth experiment, the pre-trained weights trained for 3x are used.

TABLE II. MODEL STRUCTURE AND RESULTS OF ABLATION EXPERIMENTS

Exp.	DCN	PAFPN	mAP	AP_{50}	AP_S	AP_M	AP_L
1	-	-	40.4	92.3	19.2	34.7	46.8
2	√	-	41.3	92.9	19.1	36.1	47.5
3	-	√	41.2	93.7	19.7	36.5	47.0
4	√	√	42.0	93.1	20.0	37.2	47.9
5	√	√	43.0	94.2	22.5	37.9	49.2

As can be seen from Table II, under the premise that the IoU ranges from 0.5 to 0.95 with an interval of 0.05, in Experiment 1, the original SOLO V2 model can achieve an accuracy of 40.4% in the segmentation task of insulator shed skirts.

After the DCN structure is introduced in Experiment 2, the mAP increases by 0.9%. This comparative experiment shows that the deformable convolution enables the network to better extract the local feature information of the target, improves the feature representation ability of the model, and enhances the segmentation accuracy. The introduction of deformable convolution significantly improves the accuracy for medium and large-sized shed skirts, but has little improvement effect on small shed skirts. In Experiment 3, the improvement of introducing the PAFPN method alone, compared with Experiment 1, is that because a bottom-up path is added to transfer the detailed information and spatial information in the low-level feature maps to the high-level feature maps, the feature maps are fused with richer feature information, resulting in a better segmentation effect. The mAP increases by 0.8%. In Experiment 4, by combining the two methods with the basic method, the mAP is increased by 1.6% and can reach 42.0%, which proves the feasibility of this improvement scheme. In Experiment 5, the model weights of SOLO V2 trained for 36 epochs on the COCO dataset are used as the pre-trained weights of the improved model to initialize the model, and the mAP is increased by 2.6% compared with Experiment 1.

Fig. 5 and Fig. 6 visualize the segmentation results of the ablation experiments on the insulator dataset. These two figures respectively show two types of insulators. Fig. 5 depicts ceramic insulators, and Fig. 6 shows composite insulators.

Fig. 5. Comparison of the segmentation results of ceramic insulators from ablation experiments.

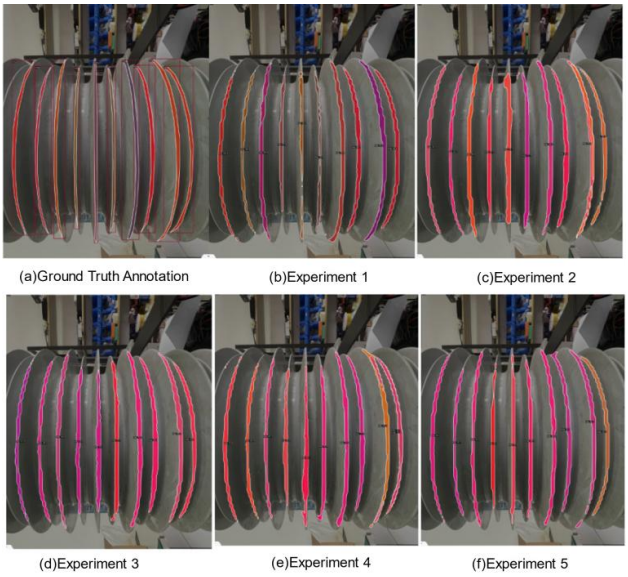

Fig. 6. Comparison of the segmentation results of composite insulators from ablation experiments.

From these two figures, it can be observed that the basic model has the problem of missed segmentation when segmenting the insulator sheds. In the first image, a complete shed instance was not segmented out. In the second image, the segmentation of two middle sheds was poor, with some parts of the sheds being missed. In Experiment 2, where DCN was introduced, this problem was alleviated, indicating that deformable convolutions can better extract feature information, but the effect was not very satisfactory. In Experiment 3, by using the PAFPN to fuse feature maps from different paths and transfer semantic and positional information of features, the segmentation accuracy was improved, and the segmentation accuracy of the sheds was significantly enhanced. In Experiment 4, where the above two methods were combined, as well as in Experiment 5, the segmentation results were the best, and the edges of the sheds were the smoothest, indicating that good pre-trained weights can significantly improve the model's accuracy.

In order to prove the effectiveness of the experiment, comparative experiments before and after the model improvement were conducted on two public datasets, CVPPP and COCO. The experimental results are shown in Table III and Table IV. Due to the more complex background of the COCO dataset and the significant differences in the number and size of targets compared to the CVPPP dataset, the mAP of the mask is lower than that of the CVPPP dataset. However, the improved model outperforms SOLO V2 in instance segmentation on both datasets. On the CVPPP dataset, the average accuracy of the improved model increased by 1.5%, and the AP50 increased by 2%. The segmentation accuracy of leaves of different sizes has been improved. On the COCO dataset, the mAP increased by 1.3%.

The comparison of the prediction results of the model SOLO V2 before and after the improvement on the CVPPP leaf segmentation dataset is shown in Fig. 7. The different rows in the image represent the segmentation results of leaves at different scales. The first column in the image is the ground truth annotation image, and the second and third columns are the segmentation results before and after the model improvement, respectively. It can be found that the segmentation effect of the improved model on both large-sized and small-sized leaves has been significantly improved, and it can segment the leaves more accurately. The improvement effect of the improved model on small leaves and leaves in densely distributed areas in the middle is particularly obvious, and there are significant improvements in the cases of missed segmentation and segmentation errors. Fig. 8 shows the segmentation results of the model before and after the improvement on the COCO dataset, and it can also be seen that the segmentation effect of the improved model is better than that of the basic model.

TABLE III. SEGMENTATION RESULTS FOR THE CVPPP DATASET

Model	mAP	AP_{50}	AP_S	AP_M	AP_L
SOLO V2	62.2	83.3	36.3	82.3	83.7
Improved	63.7	85.3	39.3	82.9	86.3

TABLE IV. SEGMENTATION RESULTS FOR THE COCO DATASET

Model	mAP	AP_{50}	AP_S	AP_M	AP_L
SOLO V2	34.8	54.9	13.4	37.8	53.7
Improved	36.1	56.0	14.8	39.1	56.1

Fig. 7. Segmentation results of different size plants for the CVPPP dataset.

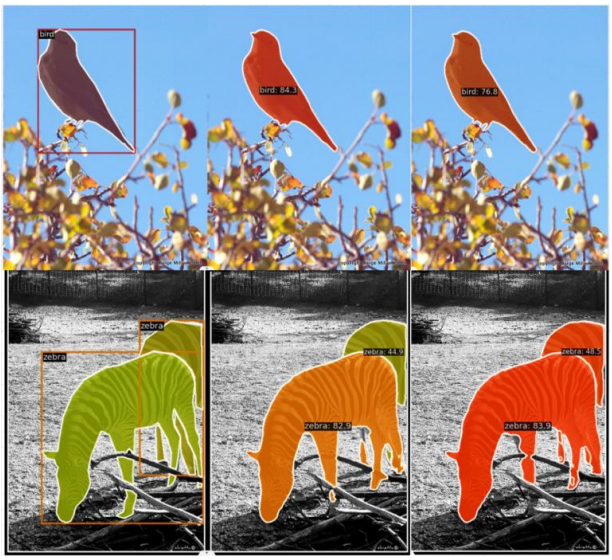

Fig. 8. Plot of segmentation results for the COCO dataset.

E. Comparison of Different Models

To verify the segmentation effect of the DPA-SOLOV2 method proposed in this section, this study conducted comparative tests on instance segmentation algorithms such as Mask RCNN, CondInst, Yolact++, SOLO V1, SOLO V2, and the R2SC-Yolact++[22]. The experimental results are shown in Table V. The table lists the mAP of each model in the insulator dataset. Compared with other models, DPA-SOLOV2 has the highest average accuracy and the best segmentation effect for the insulator shed skirts.

The experimental results show that CondInst has the worst segmentation performance. Mask RCNN and SOLO perform slightly worse on the insulator dataset, and their mAP is slightly lower than that of the baseline model. SOLO V2 and Yolact++ have better segmentation results. SOLO V2 can dynamically segment each instance in an image without the need to rely on bounding box detection. It differentiates the masks of different instances according to the size and position information of the instances. The mAP of the baseline model SOLO V2 is significantly higher than that of other methods. The improved model has addressed the issues existing in the baseline model, with its accuracy increased by 1.5% compared to the baseline model. The segmentation effect of the improved model based on SOLO V2 is better than that of R2SC-Yoolact++. The average precision of DPA-SOLOV2 can reach 43%, which is 1.3% higher than that of R2SC-Yolact++, and it has the best segmentation performance. The reasons can be summarized as: 1) The introduced DCN can adaptively adjust the shape of the convolution kernel, enabling better extraction of feature information for targets of different scales and shapes. 2) PAFPN transfers detail information from shallow features to high-level feature maps, while better conveying the positional information of features.

TABLE V. COMPARISON OF THE MODEL SEGMENTATION RESULTS

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
R2SC-Yolact++	Resnet50	41.7	82.3	24.5	34.3	49.9
Ours	Resnet50	43.0	94.2	22.5	37.9	49.2

Fig. 9 shows the comparative segmentation results of the insulator sheds by DPA-SOLOV2 and its best competitor, R2SC-Yolact++. The results indicate that the R2SC-Yolact++ has the problem of missed detections in images with a large number of sheds, and its segmentation ability for sheds at a large angle above or below the lens is relatively poor. DPA-SOLOV2 has improved the problem of target missed detection in R2SC-Yolact++ and enhanced the target segmentation accuracy. In addition, it can be observed from the image that DPA-SOLOV2 does not segment the edges of the sheds smoothly enough. It is equivalent to sacrificing some smoothness of the segmentation edges of the images to improve the problem of missed detections in the images. Further in-depth research should be carried out on this issue in the future.

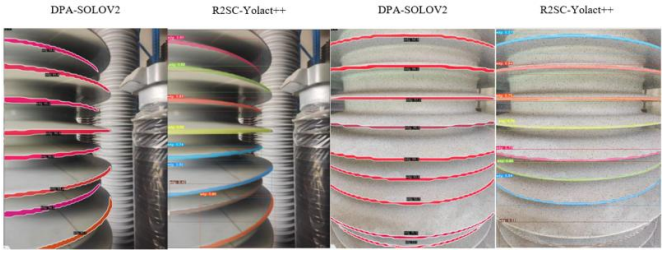

Fig. 9. Comparison of shed segmentation results of R2SC-Yolact++ and DPA-SOLOV2.

F. Discussion

In this study, the segmentation performance of the model was verified and compared on two public datasets and a self-made insulator dataset. To more conveniently compare and demonstrate the segmentation effects of the analysis model on different datasets, this study uniformly used the COCO evaluation metrics. The performance of the SOLO V2 model after introducing the deformable convolution network and PAFPN was compared and analyzed on the insulator dataset, and the performance improvement of the model before and after the improvement was verified on the two public datasets. The experimental results of the three datasets also show that the model has achieved greater improvements in the segmentation of large objects. This study can accurately locate targets and achieve pixel level segmentation to distinguish instances of the same category. It can be applied to industrial inspection (such as positioning insulator shed), healthcare (such as segmenting tumor cells), autonomous driving (such as identifying road targets) and other fields to promote intelligent development in multiple domains. However, DPA-SOLOV2 focuses on solving the problems of missed detections and false detections in instance segmentation, without paying attention to whether the edges of instance segmentation are smooth. Subsequent research will further optimize this model.

V. CONCLUSION

In this study, SOLO V2 is used as the baseline model, aiming to improve the accuracy of the model by addressing the issues of segmentation errors, missed segmentations, and low edge segmentation accuracy. First, we introduce a deformable convolution structure into ResNet50, enabling the network to better extract local feature information of targets with different scales and shapes, thereby enhancing model performance. Second, we replace FPN with the PAFPN feature fusion method to strengthen feature information fusion. PAFPN adds a bottom-up path to transmit feature spatial information, enhancing information interaction between features and allowing the model to locate targets more accurately. The model is trained and tested on the public datasets COCO and CVPPP to verify the effectiveness of the improved model. After the improvement, the mask average accuracy on the COCO dataset increases by 1.3%, and the mask average accuracy on the CVPPP dataset increases by 1.5%. The improved model is applied to the self-annotated insulator dataset to segment the umbrella skirt part of the insulators. The experimental results show that a more

accurate segmentation of the umbrella skirt of insulators is achieved.

Although the improved model has enhanced the accuracy of instance segmentation on three datasets, its edge segmentation for insulators remains relatively rough. In the future, it is possible to explore how to achieve smoother edge segmentation by preprocessing methods such as image sharpening and contrast enhancement to highlight target edges. Additionally, to further enhance the model's generalization ability, more insulator images of different types and environments should be collected and annotated in the future to enrich the insulator segmentation dataset. Given the high time cost of image annotation, subsequent research can also focus on weakly supervised instance segmentation methods.

ACKNOWLEDGMENT

This work is supported by the Youth Innovation Team Development Plan of Shandong Province Higher Education (2019KJN048).

REFERENCES

[1] Wang Z, Men S, Bai Y, et. al. “Improved Small Object Detection Algorithm CRL-YOLOv5,” Sensors (Basel), vol. 24, no. 19, p. 6437, 2024.

[2] Liu J, Guan W. A summary of traffic flow forecasting methods[J]. Journal of highway and transportation research and development, 2004, 21(3): 82-85.

[3] Sun C, Chen Y, Qiu X, Li R, You L. “MRD-YOLO: A Multispectral Object Detection Algorithm for Complex Road Scenes,” Sensors (Basel), vol. 24, no. 10, p. 3222, 2024

[4] REN S Q, HE K M, GIRSHICK R, et al. “Faster R-CNN: towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[5] REN S Q, HE K M, GIRSHICK R, et al. “Faster R-CNN: towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[6] REN S Q, HE K M, GIRSHICK R, et al. “Faster R-CNN: towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017.

[7] Bolya D, Zhou C, Xiao F, Lee Y J, “YOLACT++:better real-time instance segmentation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 2, pp. 1108-1121, 2020.

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

[9] 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.

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

[11] 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.

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

[13] LI Y, QI H Z, DAI J F, et al. “Fully convolutional instance-aware semantic segmentation,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, pp. 4438-4446 , 2017.

[14] CHEN H, SUN K Y, TIAN Z, et al. “BlendMask: top-down meets bottom-up for instance segmentation,” Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, pp. 8570-8578, 2020.

[15] WU X R, QIU T T, WANG Y N. “Multi-object detection and segmentation for traffic scene based on improved Mask R-CNN,” Chinese Journal of Scientific Instrument, 2021.

[16] YAN T R, MA X J, RAO Y L, et al. “Rebar size detection algorithm for intelligent construction supervision based on improved Mask R-CNN,” Computer Engineering, vol. 47, no. 9, pp. 274-281, 2021

[17] 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.

[18] 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, p. 126812, 2023.

[19] LIU W B, YE T, LI Q. “Tomato leaf disease detection method based on improved SOLO v2,” Transactions of the Chinese Society for Agricultural Machinery, vol. 52, no. 8, pp. 213-220, 2021.

[20] D. P. Zolg et al., “MSIS: Multispectral Instance Segmentation Method for Power Equipment ,” Comput Intell Neurosci, vol. 4, p. 2864717, 2022.

[21] Yu Z, Liu L, Jiao H, Chen J, Chen Z, Song Z, Lin H, Tian F. “Leveraging SOLOv2 model to detect heat stress of poultry in complex environments,” Front Vet Sci. vol. 6; no. 9, pp 1062559, 2023.

[22] Liqun M, Chuang C, Haonan X, Xuanxuan F, Zhijian Q and Chongguang R, “Instance Segmentation Method based on R2SC-Yolact++,” International Journal of Advanced Computer Science and Applications(IJACSA), vol. 14, no. 10, 2023.