ABSTRACT Two-stage instance segmentation method outperforms the one-stage counterpart on complex occasions. However, we found that the RoIAlign operation identifies the feature map to smaller size, and the convolution or up-sampling causes the loss of detailed information. All these make it difficult to achieve precise segmentation. To circumvent the issue, we propose a simple and efficient anchor-free model for instance segmentation. We name it as CSAS because it combines the detection-based and segmentation-based idea. The CSAS adopts the two-stage paradigm, which mainly includes detection and segmentation. The box head not only considers the location accuracy into confidence score but calculates the IoU loss of regression, which leads to a gain of 1.5%. The mask head adopts the multi-task learning to accomplish precise segmentation, and it grows 1.7 points. Using the ResNet-50-FPN, a single CSAS obtains 1.6% improvement over the Mask R-CNN. Our result demonstrates that CSAS is capable of gaining the complete mask of instance. We conclude that the detailed feature information is essential for precise segmentation, the idea is available for other segmentation tasks.

INDEX TERMS Instance segmentation, multi-task learning, anchor-free model, instance boundary.

I. INTRODUCTION
Instance segmentation is a fundamental but challenging task in computer vision, whose goal is to classify pixels at the instance level and label different instance regions. Although the research of instance segmentation has been rapidly progressing in recent years, realizing precise segmentation is still challenging due to three factors: imprecise location, background confusion, and feature loss.

Non-cascade instance segmentation methods are mainly divided into two categories: two-stage method and one-stage one. While the one-stage models achieve the real-time segmentation, the performance is worse than the two-stage counterparts. Hence, the top performing segmenters often follow the two-stage paradigm.

The most representative two-stage method, Mask R-CNN [1] employs the region proposal network (RPN) [2] to distinguish objects, and then a semantic segmentation branch predicts binary mask inside bounding boxes (Boxes).

In the detection branch of Mask R-CNN, the RPN [2] is little enough to fit various shapes of objects as an anchor-based detector. The excessive number of anchors causes the imbalance of positive and negative samples, which consumes higher computation cost. The complex anchor frame reduces the speed of network segmentation.

Recent advances in one-stage object detection prove that the one-stage detectors outperform their two-stage counterparts. CenterMask [3] and PolarMask [4] accomplish real-time segmentation because they are built on the one-stage detector FCOS [5].

The anchor-free FCOS breaks the limit of anchor-based detectors and avoids complicated computation about anchors. We also employ it to locate objects. Furthermore, we take location accuracy into consideration when ranking bounding boxes in the FCOS head. The weight of bounding boxes with poor regression performance is decreased like Mask Scoring R-CNN [6].
In the segmentation branch of Mask R-CNN, the RoIAlign [1] operation is necessary to get RoI-wise feature, whereas it causes the misalignment between feature map and object. Moreover, the cropping feature map is identified to smaller size (e.g., $7 \times 7$ or $14 \times 14$) which further loses details. These make it difficult to obtain precise mask of instance, especially the complete boundary of instance.

A line of researches [7], [8], [9] take a segmentation-based strategy to address the issue. These methods generate dense embedding features of pixels firstly, and then group them to generate final mask by post-processing operations. Although these models have poor generalization, the position information and local coherence are well retained. Therefore, we add a supervised semantic segmentation branch to settle the misalignment and weaken the confusion of messy background. In addition, we also notice that a single semantic branch is poor to process the counters of instances when they overlap each other. An excessive edge segmentation branch is applied to solve the mutual overlapping.

When extracting features of image, FPN [10] loses certain spatial information as a traditional component. To overcome the obstacle, we explore a powerful structure named Dilated Feature Extraction Pyramid (DFEP).

In this paper, we propose a model CSAS for instance segmentation combining the bottom-up and top-down idea. It adds supervised edge and semantic supplements for instance segmentation. We reinforce the information flow and strengthen the relationship among multiple tasks. All these obtain the precise mask, including the boundary of object.

Our main contribution can be summarized as follows.

1) We correct the weak correlation between location accuracy and Box confidence in the FCOS head. Compared with initial FCOS, the Boxes with poor regression are filtered. The optimization improves the accuracy of detection.

2) We propose the instance segmentation model which introduces the supervised edge head and semantic head to achieve precise segmentation. Since the idea of our work is easily implemented, it can be extended to other principles for instance-level recognition tasks.

3) We explore a powerful feature extraction module to provide the detailed information from the reduction of information loss and the improvement of features delivery. It is available to embed into other existing models.

4) We strengthen the relationship among multiple tasks. With the guide of attention mechanism, the context information is enhanced and brings vital performance development.

II. RELATED WORK

A. INSTANCE SEGMENTATION

The non-cascade models of instance segmentation include two categories: two-stage methods and one-stage methods.

1) TWO-STAGE METHODS

Two-stage methods contain the detection-based (top-down) paradigms and the segmentation-based (bottom-up) ones.

Detection-based methods follow “detect then segment” strategy. The classical algorithm is Mask R-CNN [1] that detects object and performs binary classification within Boxes. PANet [12] proposes a Bottom-up Path Augmentation to decrease the loss of information and an adaptive feature pooling operation to integrate multi-level features. Mask Scoring R-CNN [6] contributes a scoring mechanism to correct the deviation between mask quality and mask score.

Segmentation-based methods adopt “segment then classify” strategy. SGN [7] that is inspired by watershed algorithm assembles results of three sub-tasks to generate the final mask. Brabandere et al. [8] proposes a discriminant loss function which gathers pixels belonging to the same instance and scatters pixels of different instances. SSAP [9] calculates the probability of two pixels belonging to the same instance. It combines semantic segmentation and affinity pyramid joint learning to generate multi-scale prediction, and then sequentially produces the mask through cascade graph partition.

2) ONE-STAGE METHODS

Recent studies on the one-stage methods [13], [14], [15], [16], [17], [18] attract more attention because of the development of one-stage detectors, some still adopt the paradigm of two-stage but replace the RPN with one-stage alternatives. CenterMask [3] adds the spatial attention-guided mask (SAG-Mask) branch to anchor-free objector FCOS [5]. YOLACT [16] achieves real-time segmentation by linearly combining the k prototype masks and per-instance mask coefficients. BlendMask [18] designs a blender to further fuse the global semantic information and detailed location information. PolarMask [4] represents the instance with polar coordinates firstly. It transforms the instance segmentation into the classification of instance center and the dense distance regression. Deep-snake [19] adjusts the counter of instance step by step on the basis of traditional snake and Curve-GCN [20] algorithm.

B. METHODS USING SEMANTIC SEGMENTATION

Semantic segmentation well maintains the details of image and keeps local-coherence. Certain instance segmentation models introduce it as the supplement. HTC [21] introduces a semantic segmentation branch to distinguish objects from clutter background. MaskLab [22] employs a semantic branch to discriminate semantic categories of pixels, and a direction branch is used to predict the distance of each pixel to the corresponding instance center.

C. ATTENTION MECHANISM

Attention modules effectively improve the performance of a deep convolutional neural network. The Squeeze-and-Excitation(SE) [23] attention captures channel correlations by selectively modulating the scale of channel.
The CBAM [24] enriches the attention map by adding max pooled feature for the channel attention. The Non-Local block [26] captures the long-range dependency using the non-local operation to build a dense spatial feature map.

III. METHOD
In this section, we present the framework of CSAS in Fig 1. It attaches a novel fused segmentation branch to one-stage FCOS [5] in the same vein as Mask R-CNN. We optimize the original FCOS to accurately locate objects. Based on the powerful DFEP, the multimaps with rich context information are delivered into three network modules (i.e., the semantic head, the mask head, and the edge head) to perform precise instance segmentation. In addition, the attention mechanism is utilized to highlight the informational pixels and suppress the noise when producing the feature maps. The details of above components are stated as follows.

A. DILATED FEATURE EXTRACTION PYRAMID
The high-level layers of pyramid with wider receptive filed capture overall semantic information such as pose, while the low-level layers preserve more detailed information such as location. To improve the expression of features and get detailed information, we propose the DFEP as shown in Fig 2. It enhances the feature information from two aspects: developing the delivery efficiency and enriching the context information. The Boxes with larger scales have to be assigned to a higher level and vice versa. To balance the performance and overhead, we adopt an Adaptive RoI Assignment Function which considers the ratio of input and RoI area.

1) DELIVERY EFFICIENCY IMPROVEMENT
The FPN [10] operates the up-sampling and feature fusion iteratively, certain significant information loses during the convolution or deconvolution. A novel attention mechanism named ESCA makes the network focus on the important information and decrease the loss of spatial information. The ESCA is composed of two parallel branches as shown in Fig 3. In the channel branch, the feature map $R^{256 \times H \times W}$ is processed with a global average pooling to obtain dependency along the channel axis, and a sigmoid activation function is utilized to achieve the adaptive selection of channels. In the spatial branch, we add a convolutional layer to weigh the importance of different areas. We sum the outputs of two branches to feed into next tasks.

2) ENRICH THE CONTEXT INFORMATION
Because the different scales of RoIs are mapped into appropriate levels, there is no information interaction among multiple levels in feature pyramid. The single feature map is bound to contain more information for next tasks. We further expand the receptive filed of the smallest feature map generated by backbone, as denoted in Fig 4, a Dilated Encoder Module(DEM) [27] module is followed.

B. ONE-STAGE DETECTOR FCOS
In the first phase, an efficient and simple detector is demanded to locate and distinguish objects. FCOS solves objects detection in a per-pixel prediction fashion, which predicts a 4D vector plus a class label at each spatial location on a level of feature map. These pixels falling on the GT bounding box are regarded as positive samples, which avoids hyper-parameters to label box in many anchor-based detectors. As an anchor-free detector, it breaks the limit of pre-define anchors and avoids the complicated computation related to anchor boxes. Because of the good performance and efficiency, we employ it to yield Boxes.

To improve the detection performance, we optimize FCOS as follows.

1) We calculate the IoU loss of the predicted Box and GT Box, which further enforces the regression branch to correct the location.
2) A IoU-Aware [28] branch is introduced to solve the correlation between classification score and location accuracy. It enhances the effect of location accuracy on final Box confidence.

The optimized FCOS achieves the first step to obtain precise mask of instance with negligible computational overhead.

C. FUSED SEGMENTATION BRANCH
In the second phase, the feature maps corresponding to RoIs are delivered into the fused segmentation branch to predict binary instance mask. Different from the fully convolutional network (FCN) [29] branch of Mask R-CNN, we introduce two supervised branches to get complete boundary of instance and solve the confusion of clutter background, as shown in Fig 5.

1) CHANNEL ATTENTION AND EDGE ATTENTION
The attention mechanism as a necessary trick is applied to CSAS, we employ the channel attention in mask segmentation branch and edge attention in edge segmentation branch.

a: EDGE ATTENTION
The edge branch is based on the position attention module of DANet [30] to capture the long-range contextual information in spatial dimension. The structure of position attention module is in Fig 6.

As illustrated above, the given feature map $A \in R^{256 \times H \times W}$ is fed into three convolutional layers to generate three maps $B, C, D$ where $B, C \in R^{32 \times H \times W}$ and $D \in R^{256 \times H \times W}$. Then we shape them to $R^{C \times N}$ where $N = H \times W$ is the number of pixels. After performing a matrix multiplication between the transpose of C and B, the softmax layer is followed to calculate the spatial map $S \in R^{N \times N}$.

$$s_{ji} = \frac{\exp(B_j C_i)}{\sum_{i=1}^{N} \exp(B_i C_i)}$$ (1)
FIGURE 1. Overall framework of CSAS for instance segmentation. The backbone and DFEP extract features of images, the one-stage FCOS is employed to distinguish objects, and the fused segmentation branch generates precise binary instance mask within Boxes.

FIGURE 2. Architecture of Dilated Feature Extraction Pyramid. The output channel of backbone is identified to 256, which reduces the computational overhead before the further processing. The $B_i$ and $D_i$ imply feature maps generated by backbone and DFEP, respectively.

FIGURE 3. Structure of ESCA. The $\otimes$ is element-wise multiplication operator, and the $\oplus$ is element-wise summation descriptor.

where $s_{ji}$ measures the $i^{th}$ position's impact on the $j^{th}$ position. The more similar feature representations of the two positions contribute to greater correlation between them. We multiply D with the transpose of S and reshape the result to $R^{256 \times H \times W}$. The element-wise summation is performed with D to obtain final output E.

$$E_j = \sum_{i=1}^{N} (S_{ji}D_j) + A_j$$

(2)

b: CHANNEL ATTENTION
We employ the variant of SE [23] block as the channel attention module as shown in Fig 7. It is processed by a convolutional layer (kernel sizes are $3 \times 3$ and $5 \times 5$) to expand the receptive filed firstly. We employ one connected layer to reduce the computational overhead and decrease the loss of information caused by dimension reduction. The channel attention guides the network to achieve better discrimination in the background and foreground.

2) SUPERVISED EDGE SEGMENTATION BRANCH
The edge branch encodes a wider range of contextual information into the local features and enhances the representation capability. We use a specific convolution to calculate the GT boundary corresponding to GT mask, the padding operation keeps their sizes same. The pixels falling on the contour are marked with one and others are zero as the background, the kernel weight of convolution is defined as a $3 \times 3$ matrix:

$$matrix\_size = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 8 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Let F denote the intermediate output of convolutional layer, the GT boundary $B$ is defined as.

$$B(i,j) = \begin{cases} 1, & \text{if } F(i,j) \in (7, 12) \\ 0, & \text{otherwise} \end{cases}$$

(3)

What's more, we notice that the quantities of positive and negative samples are unbalanced. The focal loss [31] function is used to calculate the loss of edge segmentation branch.

3) SUPERVISED SEMANTIC SEGMENTATION BRANCH
As the auxiliary branch, the semantic branch of CSAS is a fully convolutional neural network consisting of four layers, the highest resolution feature map of DFEP serves as the input of semantic branch. We add a convolutional layer to predict the probability of each pixel belonging to the foreground and another convolutional layer to integrate the semantic feature. The binary cross-entropy is utilized to calculate the loss of semantic segmentation branch.
Y. Yang et al.: Combine Supervised Edge and Semantic Supplement for Instance Segmentation

**Figure 5.** Overall architecture of fused segmentation branch composed of three heads, the Edge Head (edge branch), Mask Head (mask branch), and Semantic Head (semantic branch). The orange region denotes the area of a bounding box.

**Figure 6.** Architecture of position attention module of DANet.

**Figure 7.** Architecture of channel attention.

The semantic branch provides a pixel-wise prediction of the whole image. It implicitly helps to remain the boundary of objects.

4) **FEATURE AUGMENTATION MODULE**

Although some models introduce the independent semantic branch to improve the performance, some details information is still lost. These incorporate the semantic feature before the instance-wise pooling operation. In our model, we propose a FAM module to fuse the instance-wise feature and corresponding properties as shown in Fig 8.

We first concatenate the RoI-wise semantic feature, RoI-wise semantic segmentation, and instance feature. Then the convolutional operation with different dilation is adopted into FAM to capture rich semantic information. The output is named fused instance feature and used to predict final instance mask.

**IV. RESULTS**

**A. DATASETS AND EVALUATION METRIC**

1) **DATASETS**

All experiments are performed on the challenging COCO [35] dataset, which contains about 118k images with responding annotations as the training set and 5k held-out images with annotations as the validation set. Typical instance annotations are used to the supervised detection branch and mask branch.

2) **EVALUATION METRICS**

We use the standard mask AP metric that averages APs across IoU thresholds from 0.5 to 0.95 with an interval of 0.05 as the evaluation metric. Both box AP ($AP_b$) and mask AP ($AP^*$) are evaluated. We report ($AP$ at different IoU thresholds=$0.5$ and $=0.75$). We also report ($AP$ at different scales: $AP^*_S\ ,\ AP^*_M\ ,\ AP^*_L$) for the mask AP.

**B. IMPLEMENT DETAILS**

We follow the hyper-parameters of FCOS except for positive score threshold 0.03 instead of 0.05, because it lacks positive RoI samples in initial training time. We use mmdetection [36] as the code base and re-implement other previous architectures for a fair comparison.

1) **TRAINING**

We set the “max_per_image” parameter of FCOS detector to 100 during training, then the highest-confidence Boxes are fed into the mask segmentation branch. We define the loss of the fused segmentation branch $L_M$ as:

$$L_M = L_{i\_mask} + \lambda_1 L_{s\_mask} + \lambda_2 L_{e\_mask}$$

(4)

where the semantic segmentation loss $L_{s\_mask}$, the instance segmentation loss $L_{i\_mask}$, and the edge segmentation
loss $L_{\text{ce,mask}}$. We set $\lambda_1 = \lambda_2 = 0.5$ to balance the contributions of different segmentation heads. We define the loss of the supervised detection branch $L_F$ as:

$$L_F = L_{\text{cls}} + L_{\text{cen}} + L_{\text{Box}} + \lambda_3 L_{\text{IoU}} + \lambda_4 L_{\text{IoU,aware}}$$ (5)

where the classification loss $L_{\text{cls}}$, the centerness loss $L_{\text{cen}}$, the regression loss $L_{\text{Box}}$, the calculating IoU loss between the detected Box and GT Box $L_{\text{IoU}}$, the predicted IoU loss $L_{\text{IoU,aware}}$. We set $\lambda_3 = 0.25$, $\lambda_4 = 0.5$ to balance the contributions of different detected sub-tasks.

We train our model on a single RTX3080 GPU setting the batch size to 1 for 20 epoches with an initial learning rate of 0.0025, and decrease it by 0.001 after 16 and 19 epoch, respectively. We use SGD as the optimizer with a weight decay of 0.0001 and a momentum of 0.9. The long edge and short edge of each image are resized to 1333 and 800 without changing the aspect ratio. Moreover, we do not apply these tricks(e.g., extending training time, the multi-scale training).

2) INFERENCE
In the regression phase, the confidence of bounding boxes is obtained from the predicted IoU score multiplied by the classification score. We filter out these Boxes with confidence lower than 0.001 and use non-maximum suppression (IoU threshold = 0.5) to remove the duplicated Boxes.

### TABLE 1. Comparison with Mask R-CNN in different backbones and different necks.

| Method     | Backbone      | $AP_50$ | $AP^{*}_{50}$ | $AP_75$ | $AP^{*}_{75}$ |
|------------|---------------|---------|---------------|---------|---------------|
| Mask R-CNN | R-50+FPN      | 37.7    | 34.1          | 54.8    | 36.5          |
| CSAS       | R-50+DPEP     | 38.9    | 35.7          | 56.5    | 38.3          |
| CSAS       | V-39+FPN      | 40.2    | 36.2          | 57.6    | 38.9          |
| CSAS       | MobileNetV2 [39]+FPN | 31.6    | 28.6          | 46.0    | 30.6          |
| CSAS       | HRNet W18 [40]+HRFPN | 36.9    | 33.2          | 52.8    | 35.4          |
| Mask R-CNN | R-101+FPN     | 39.6    | 35.8          | 56.0    | 38.8          |
| CSAS       | R-101+DPEP    | 40.7    | 36.6          | 56.8    | 39.7          |
| CSAS       | V-57+FPN      | 41.2    | 37.1          | 58.0    | 39.7          |
| CSAS       | HRNet W32 [40]+HRFPN | 40.7    | 35.5          | 55.8    | 38.1          |
| Mask R-CNN | X-101+FPN     | 41.1    | 37.0          | 58.6    | 39.8          |
| CSAS       | HRNet W40 [40]+HRFPN | 41.5    | 36.3          | 56.9    | 39.1          |
| CSAS       | X-101+DPEP    | 41.8    | 37.6          | 59.1    | 40.1          |
| CSAS       | V-99+FPN      | 42.2    | 37.8          | 59.5    | 40.2          |

### TABLE 2. Comparison with other segmentation approaches.

| Method     | Backbone      | Epoch | $AP_50$ | $AP^{*}_{50}$ | $AP_75$ | $AP^{*}_{75}$ |
|------------|---------------|-------|---------|---------------|---------|---------------|
| BlendMask  | R-50+FPN      | 36    | 37.0    | 58.9          | 39.7    |
| SOLovl     | R-50+FPN      | 72    | 36.8    | 58.6          | 39.0    |
| MaskLab+   | R-101(JET)    | -     | -       | 61.1          | 40.4    |
| PANet      | R-50+FPN      | 24    | 36.6    | 58.0          | 39.3    |
| Mask R-CNN | R-101+FPN     | 24    | 41.6    | 37.4          | 59.5    | 40.0          |
| Mask Scoring R-CNN | R-101+FPN   | 24    | 41.6    | 38.3          | 58.5    | 41.5          |
| Cascade Mask R-CNN | R-50+FPN    | 20    | 42.8    | 37.0          | 58.6    | 39.9          |
| CSAS       | V-57+DPEP     | 20    | 41.6    | 37.5          | 59.1    | 40.4          |
| CSAS       | R-101+DPEP    | 20    | 41.2    | 37.3          | 58.9    | 39.9          |

### FIGURE 8. Architecture of Feature Augmentation Module (FAM). The orange region denotes a bounding box, and the parameter ‘d’ implies the dilation in the convolution.

### FIGURE 9. COCO val2017 example results from CSAS.

### C. MAIN RESULTS

We conduct experiments on COCO val2017 to show the performance with different backbones and different necks. The experimental result shows that CSAS achieves consistent improvements over Mask R-CNN as summarized in Table1. It achieves a gain of 1.6%, 0.8% and 0.6% for ResNet-50 [37], ResNet-101 [37], and ResNeXt-101 [38], respectively. We take some examples of segmentation result in Fig 9.

We compare CSAS with other instance segmentation approaches on COCO dataset in Table 2. Without bells and whistles, the CSAS achieves 1.7 points AP improvement over the Mask R-CNN baseline using ResNet-50-FPN as the backbone network. From Fig 10, our model achieves precise segmentation on complex occasions. Especially, it maintains the instance boundary.

### D. ABLATION STUDY

We conduct ablation studies to verify the contribution of each component we introduce for the CSAS. All studies are conducted on COCO val2017 using ResNet-50-FPN as the backbone. To prove the validity of single component, we only replace RPN with FCOS in the component-wise analysis.
1) COMPONENT-WISE ANALYSIS
We explore the effectiveness of the main components in our framework. Because our model is built on FCOS, we first study the optimized detector after introducing the IoU-loss or IoU-aware. To highlight the effectiveness of mask branch proposed in our model, we compare it with original segmentation branch of Mask R-CNN. From Table 3, we can learn that IoU-loss calculation and IoU-aware prediction improve the $AP_b$ by 1.5%, the mask branch of our model contributes to a 1.7% improvement, and the DFEP leads to a gain of 0.3%.

2) EFFECTIVENESS OF DFEP
To investigate the effect of DFEP, we compare it with original FPN and PAFPN [12]. As shown in Table 4, the DFEP outperforms them. We also study that the dilated encoder introduces more semantic information and the ESCA suppresses noise of feature information, with an improvement of 0.2% and 0.2%, respectively.

3) EFFECTIVENESS OF SUPERVISED SEMANTIC BRANCH
The semantic results work as supplements for FAM module and weaken the confusion of background. We exploit that contextual feature brings 0.6% improvement after introducing the semantic branch. In addition, we further study the influence of the semantic input in Table 5. Although more excellent semantic segmentation algorithms (e.g., DeepLab V3+ [41]) bring larger development, we adopt the traditional FCN to maintain the simplicity of CSAS.

4) EFFECTIVENESS OF SUPERVISED EDGE BRANCH
In the above section, we explicitly design an edge branch to predict counter of instance. From Table 6, the edge branch brings 1.0% improvement. We find that introducing the boundary branch greatly boosts the ability to realize precise segmentation. Without an edge branch, the segmentation of overlapping objects is limited because the main effect of semantic branch is to decrease the confusion of background.

5) EFFECTIVENESS OF FEATURE AUGMENTATION MODULE
The multi-task learning is known to be beneficial. Hence, the RoI-wise semantic feature, RoI-wise semantic mask, and RoI-wise instance feature serve as the supplements for the fused instance feature. We explore the necessity of feature fusion in Table 7, the fusion operation achieves 0.4% improvement. It indicates that the information flow brings rich enough information to obtain precise instance mask.

V. CONCLUSION
In this work, we propose an anchor-free CSAS which adopts two-stage paradigm for instance segmentation. We design
DFEP and FAM to enhance information propagation in pipelines. The detection branch corrects the association between classification score and location accuracy. The segmentation branch adopts multi-task learning fashion to form a positive feedback among sub-task branches. A single CSAS outperforms 1.6% over Mask R-CNN and alleviates the indistinct segmentation on instance boundary. Comprehensive experiments and ablation studies demonstrate the effectiveness of each module, which suggests our model is reasonable.

According to the visualization of CSAS, the segmentation of huddled instances is unsatisfactory. The main cause is that occlusion affects the feature extraction of adjacent objects. Although we employed the semantic head and edge head to improve the quality of mask, the coarse mask directly works as the output of CSAS. To address the issue, we will attempt to introduce a cascade structure to refine mask in the further work.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” in Proc. ICCV, Oct. 2017, pp. 2961–2969.
[2] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
[3] Y. Lee and J. Park, “CenterMask: Real-time anchor-free instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13906–13915.
[4] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, “PolarMask: Single shot instance segmentation with polar representation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 12193–12202.
[5] Z. Tian, C. Shen, H. Chen, and T. He, “FCOS: Fully convolutional one-stage object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9627–9636.
[6] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, “Mask scoring R-CNN,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6409–6418.
[7] S. Liu, J. Jia, S. Fidler, and R. Urtasun, “SGN: Sequential grouping networks for instance segmentation,” in Proc. IEEE ICCV, Venice, Italy, Oct. 2017, pp. 3496–3504.
[8] B. D. Brabandere, D. Neven, and L. Van Gool, “Semantic instance segmentation with a discriminative loss function,” 2017, arXiv:1708.02551.
[9] G. Zhao, Y. Shan, Y. Wang, X. Zhao, Y. Yu, M. Yang, and K. Huang, “SSAP: Single-shot instance segmentation with affinity pyramid,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 642–651.
[10] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2117–2125.
[11] L. He, Y.-W. Tai, and C.-K. Tang, “Deep occlusion-aware instance segmentation with overlapping BiLayers,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4019–4028.
[12] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, “Path aggregation network for instance segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8759–8768.
[13] W. Wang, Z. Xu, H. Shen, B. Cheng, and L. Yang, “CenterMask: Single shot instance segmentation with point representation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9313–9321.
[14] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, “SOLO: Segmenting objects by locations,” in Proc. ECCV Cham, Switzerland: Springer, 2020, pp. 649–665.
[15] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, “SOLOv2: Dynamic and fast instance segmentation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 17721–17732.
[16] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT: Real-time instance segmentation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9157–9166.
[17] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, “YOLACT++: Better real-time instance segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 2, pp. 1108–1121, Feb. 2022, doi: 10.1109/TPAMI.2020.3014297.
[41] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018, doi: 10.1109/TPAMI.2017.2699184.

[42] X. Yuan, A. Kortylewski, Y. Sun, and A. Yuille, “Robust instance segmentation through reasoning about multi-object occlusion,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 11141–11150.

YAKUN YANG received the B.S. degree in computer science and technology from the Henan University of Economics and Law, in 2020. He is currently pursuing the M.S. degree in science and technology with Hebei University, under the supervision of Prof. Luo. His main research interests include semantic segmentation, instance segmentation, and object detection.

During the period of pursuing B.S. degree, he obtains the scholarship many times and GPA of four years in university ranks the top 20%.

WENJIE LUO received the M.S. degree in computer science from Hebei University, Baoding, China, in 1998, and the Ph.D. degree in computer science from the School of Computer Science and Engineering, Beihang University, Beijing, China, in 2010.

She is currently working with the School of Cyber Security and Computer, Hebei University. She was the first inventor obtained one national invention patent. She has published 18 articles, which nine articles are included into EI and three articles are retrieved by ISTP. Her research interests include machine vision, software engineering, and personalized recommendation.

Prof. Luo has presided and completed one Hebei provincial project. She has participated in more than ten national, provincial, and ministerial projects. The research on the data structure of virtual ancient books project won the second prize of scientific and technological progress of the Hebei Provincial Department of Education. The course “data structure” was rated as the provincial excellent course of Hebei Province.

XUEDONG TIAN received the B.S. degree from the Department of Automation Engineering, Hebei University of Technology, China, in 1984, the M.S. degree from the Department of Electronics and Information Engineering, Hebei University, Baoding, China, in 1998, and the Ph.D. degree from the College of Physics Science and Technology, Hebei University, in 2007.

He is currently a Professor with the School of Cyber Security and Computer, Hebei University. He assumes the Director of the Chinese Information Society and the Executive Director of Hebei Artificial Intelligence Society. He has published more than 40 academic papers, including three in SCI, 14 in EI, and 12 in Chinese core journals. His research interests include pattern recognition, image processing, information retrieval, and computational linguistics.

Prof. Tian has presided and completed one National Natural Science Fund Project and three Natural Science Fund Projects of Hebei Province. He has won two third prizes for scientific and technological progress in Hebei Province and one first prize for scientific and technological achievements of universities in Hebei Province.