Supervised Edge Attention Network for Accurate Image Instance Segmentation

Xier Chen *, Yanchao Lian *, Licheng Jiao ✉[0000−−0003−3354−9617], Haoran Wang, YanJie Gao, and Shi Lingling

School of Artificial Intelligence, Xidian University
Xian, Shaanxi Province, 710071, China
{xechen, yclian}@stu.xidian.edu.cn
lchjiao@mail.xidian.edu.cn

Abstract. Effectively keeping boundary of the mask complete is important in instance segmentation. In this task, many works segment instance based on a bounding box from the box head, which means the quality of the detection also affects the completeness of the mask. To circumvent this issue, we propose a fully convolutional box head and a supervised edge attention module in mask head. The box head contains one new IoU prediction branch. It learns association between object features and detected bounding boxes to provide more accurate bounding boxes for segmentation. The edge attention module utilizes attention mechanism to highlight object and suppress background noise, and a supervised branch is devised to guide the network to focus on the edge of instances precisely. To evaluate the effectiveness, we conduct experiments on COCO dataset. Without bells and whistles, our approach achieves impressive and robust improvement compared to baseline models. Code is at https://github.com//IPIU-detection/SEANet.

Keywords: Fully convolutional box head, supervised edge attention module, IoU prediction branch, instance segmentation

1 Introduction

Instance segmentation is a hotspot task in computer vision, whose goal is to segment and classify the pixels belonging to instance objects in an image. Despite great advances have been achieved in recent years, how to keep the boundary of the instance intact still remains challenging due to three problems: imprecise bounding box, blurry boundary and cluttered overlap.

Instance segmentation methods mainly include detection-based and segmentation-based methods. Detection-based methods [8,22,9] use the detector to find and classify the possible bounding boxes of the instance firstly, and then perform the foreground and background segmentation on the pixels in these boxes. Segmentation-based methods [1,2,21] segment the whole image firstly, and then

* Two authors contribute equally to this work.

Fig. 1. Baseline Mask R-CNN (top) vs. Ours (bottom) with ResNet-101-FPN. Our approach is more adaptive for scales of objects and more sensitive on their boundaries (best viewed in color)

distinguish the object instances from it. This paper mainly discusses the former one, and analyzes problems in two phases of this kind of methods.

In the first phase, the quality of the bounding box is important as the base of segmentation. The results of the Mask R-CNN [9] are visualized for illustration. As shown in the first picture of Fig. 1, the tail of the train is not included in the bounding box, resulting in its tail boundary not being segmented. The similar situation also occurs in the second picture, and the mask at the top of the bird is affected by the imperfect bounding box. In order to improve the performance of bounding box regression, fully convolution network is proposed to break the limitation of the fixed anchor sizes and scales [25,14,28]. The technique predicts the classification and regression pixel by pixel. However, the classification score obtained from the box head shows no important interaction to the quality of the regression result.

Motivated by the considerations above, we propose a new branch named "B-IoU" based on the fully convolutional box head. It predicts the Intersection-over-Union (IoU) scores between the detected boxes and their ground-truth boxes. By multiplying the classification scores with IoU scores, we get the final score for each detected box. Through choosing the detected box with the highest score on each proposal from Region Proposal Network (RPN) [23], we expect to get the most accurate detected result for each proposal.

In the second phase, segmentation on the object boundary may fail for two reasons: (1) the information of edges of the objects is easily disturbed by background noise; (2) different instances are often close to each other or overlapped, and their features of the edge regions will interact with each other during convolution operation. In order to improve the discrimination ability, recently attention mechanisms are increasingly used in segmentation networks [6,27,16] to pay attention to the information of the objects. However, the available methods only focus on the correlation of internal features of the object and do not emphasize edge features. As shown in the last two pictures of Fig. 1, the wing and tail surface of the airplane in the third one are spindly and the colors

of them are similar with the ground and mountain. And in the last picture, the top edge of the bus almost merges with the sky and the people are overlapped, leading to the segmentation on the edge of them unsatisfactory.

To solve above-mentioned problems, we propose the edge attention modules to suppress useless information and highlight boundary features. Especially, we apply a supervision branch on the module to guide the network to learn the right information on the boundaries of the instances.

The second row of Fig. 1 shows the results of the Mask R-CNN equipped with our techniques. The results of the first two pictures show that regardless of the shapes and scales of the objects, more accurate detection bounding boxes are obtained, which has a direct effect on the performance of segmentation. Moreover, for the challenging scenes in the last two pictures, the boundaries of the instances, especially at the wing and tail surface of the airplane, the top of the bus and the boundary between person and bus, are segmented more subtly and correctly.

In this paper, we propose one new branch named "B-IoU" and supervised edge attention module for instance segmentation. Our main contributions are summarized as follows:

1. We take into consideration the quality of the bounding box in instance segmentation task. We apply fully convolutional box head and introduce a new branch name "B-IoU" to learn the IoU scores between the detected bounding boxes and their corresponding ground-truth boxes for down-weighting the low-quality detected bounding boxes with poor regression performance.
2. As the boundaries of the instances are easily overwhelmed by the background noise or other objects, we propose supervised edge attention module to suppress the noise and highlight the foreground. Especially, we design a supervised branch to guide the network to learn the boundaries of the objects.
3. Without bells and whistles, our approach consistently improves the models of Mask R-CNN series, and is no limited to these models. Since the idea of our work is easily implemented and can improve both the accuracy of detection and segmentation, it can be extended to other principles for instance-level recognition tasks.

2 Related Work

2.1 Instance segmentation

Instance segmentation is the task of labeling different instance regions in the pixel-level, which is divided into two categories: segmentation-based and detection-based.

Segmentation-based methods are to first segment and then detect. BIS [1] uses CRF [15] to find instances. SIS [2] introduces metric learning, and completes segmentation with clustering. SOIS [21] uses semi-convolutional operation for instance segmentation.

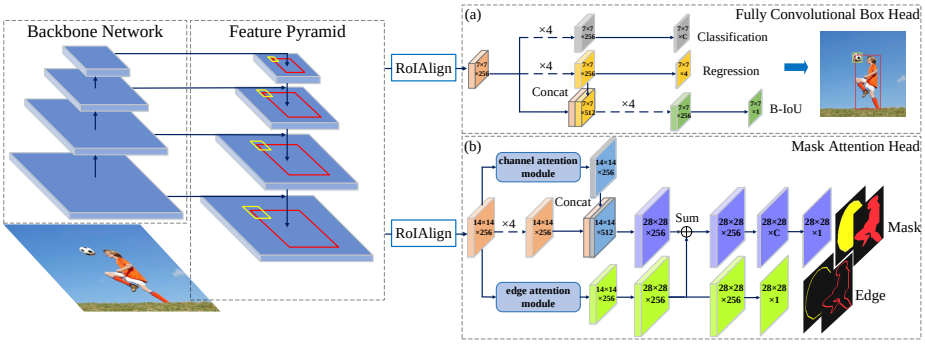

Fig. 2. The model structure of SEANet. The Feature Pyramid Network (FPN) extract the features of the input image. RoIAlign operation obtains feature map of each proposal from RPN. For each proposal, (a) fully convolutional box head and (b) mask attention head perform object detection and instance segmentation in parallel

The other idea for instance segmentation, called detection-based method, is to first detect and then segment. These methods use bounding boxes as the preliminary results, and then estimate region masks in the boxes. The Facebook Research Institute proposes Deep Mask [22], which is based on the object proposals generated by discriminative convolutional network. ISFCN [3] uses FCN [20] to generate object proposals with the help of a sliding window. He et al. proposes Mask R-CNN [9] consisting of a box head and a mask head for detection and segmentation respectively, which achieves better results in instance-level segmentation. Mask R-CNN inspires many other algorithms as a benchmark framework. Based on it, Mask Scoring R-CNN (MS R-CNN) [11] proposes a MaskIoU head to generate IoU scores to describe the qualities of masks. It combines the features of instances in mask head with the corresponding predicted masks. However, in addition to prioritizing when evaluating, the IoU scores have no substantial impact on improving the quality of the mask itself. Different from MS R-CNN, we design a B-IoU branch in box head to produce IoU scores. The IoU scores are used to obtain more precise bounding boxes so as to improve the performance of instance segmentation.

2.2 Object Detection

For the detection-based instance segmentation methods, the quality of object detection is particularly important. Faster R-CNN [23] is the base framework of Mask R-CNN, which is a widely used two-stage network. It first generates Regions of Interest (RoIs) through the RPN, and then performs further classification and regression operations on the proposals. Considering that the predicted classification confidence cannot represent the quality of the location, IoU-Net [12] proposes location confidences to predict the IoUs between detected bounding boxes and corresponding ground-truth boxes. But the IoU head in IoU-Net is

parallel with the box head, which makes no full use of the information from the regression branch. Moreover, most two-stage detectors , including IoU-Net, implement detection in the proposal level instead of the pixel level, which limits the representation of the network. FCOS [25] perform pixel-by-pixel classification and regression prediction on feature maps, and introduces a new branch called "center-ness" to suppress the low-quality bounding box. In comparison, our fully convolutional box head combines information of the high-level regression feature and the low-level RoI feature to infer IoU score for each detected box. In this way, it identifies the quality of bounding box more accurately so as to improve the quality of mask.

2.3 Attention mechanism

Attention mechanism has been widely used in the field of computer vision. SENet [10] uses squeeze module and association module to obtain the importance of each feature channel automatically. CBAM [26] uses feature-channel attention similar to SENet with additional feature-space attention. GSoP-Net [7] proposes 2D-average-pooling, which applies channel attention in the form of covariance. DANet [6] proposes position and channel attention modules, which focus on the correlation of semantic features of each position and channel. In this paper, we take advantage of the attention mechanism and propose two modules. Different from the methods above, we focus on the object boundary and design a supervised branch which utilizes explicit information to achieve it.

3 Our Approach

In this section, we first present the framework of our Supervised Edge Attention Network for Accurate Image Instance Segmentation (SEANet), then introduce the two techniques and the loss function. We take the Mask R-CNN as an example. As shown in Fig. 2, the fully convolutional box head is employed in the detection branch to obtain the appropriate bounding box for segmentation and the supervised attention module is added to the original mask head.

3.1 Fully Convolutional Box Head

Different from many box heads of the anchor-based methods, which output the class and offset result for each proposal using fully connection layers, our fully convolutional box head focuses on information of each pixel on the proposal. The structure is shown in Fig. 2 (a). For each location (x, y) on the feature map of each candidate box after RoIAlign, an 80-D classification score C, a 4-D bounding box regression $B = (l, t, r, b)$ and a 1-D box IoU score I are given by the fully convolutional box head. Here, l, t, r and b represent the distances from the location (x, y) to the left, top, right and bottom sides of the detected bounding box respectively.

Given a proposal $P = (x_0, y_0, x_1, y_1)$ from RPN module, the corresponding location of (x, y) on the input image is expressed as (X, Y):

$$X = x_0 + x\frac{W_p}{l} + \frac{W_p}{2l}, Y = y_0 + y\frac{H_p}{l} + \frac{H_p}{2l} \tag{1}$$

Here, W_p and H_p are the width and height of the proposal. (x_0, y_0) and (x_1, y_1) denote the left-top and right-bottom coordinates of the proposal respectively. And l represents the size length of the feature map of the proposal after RoIAlign ($l = 7$ in this paper). For classification, if (X, Y) is inside any ground-truth bounding box, then corresponding (x, y) will be labeled as a positive sample and given a class label c^* as same as its corresponding ground truth. Otherwise, it is considered as a negative sample and its class label $c^* = 0$ for background.

For regression, each positive sample (x, y) has a regression target as $B^* = (l^*, t^*, r^*, b^*)$, which means the distances from its corresponding location (X, Y) to the four sides of its corresponding ground-truth bounding box.

In the test phase, for each proposal from RPN module, output of our fully convolutional box head is a set of detected bounding boxes and their corresponding classification scores. However, only one final detection result is expected for each proposal region. And as mentioned in the previous section, the misalignment between classification scores and the regression qualities harm the performance of the detection. Therefore, a powerful strategy should be adopted for choosing the best regression result from detected bounding boxes produced by all the locations on the proposal.

We introduce a meaningful branch called "B-IoU" head to predict the IoU value between the predicted bounding box and its associated ground-truth bounding box at each location on each proposal. The "B-IoU" branch combines the information of the proposal feature and the high-level semantic feature from the regression branch and output a more targeted and precise evaluation criteria for each bounding box by aligning the regression quality with box IoU scores. Here, the calculation of the targeted IoU scores for training is similar to GIoU [7], which presents a more precise measure of the IoU between the detected bounding box and the corresponding ground-truth box.

During inference, the IoU scores obtained from the B-IoU branch are multiplied with the classification scores from the classification branch to get the final scores which describe the qualities of the bounding boxes precisely in both the aspects of the classification and regression. And for each proposal, only the detected bounding box with the highest final score will be chosen as the detection result of this proposal and all the detection results of all the proposals will be ranked by the final scores so that the low-quality bounding boxes will be suppressed by Non-Maximum Suppression (NMS). Moreover, more accurate scores of the detection results will benefit to the COCO evaluation process and improve the performance of instance segmentation.

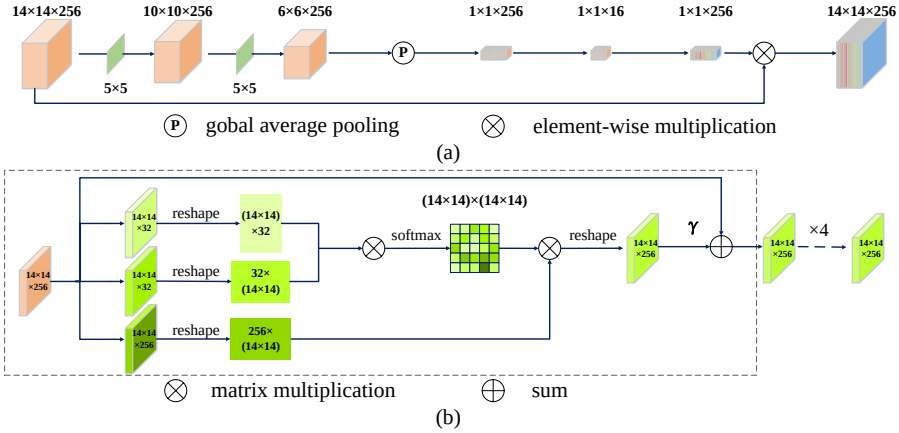

Fig. 3. Illustration of the attention modules. (a) The structure of the channel attention module. (b) The structure of the edge attention module. The part in the gridline is the position attention module in DANet

3.2 Supervised Attention Module

In order to weaken the non-object information and prevent the boundaries of the objects being blurred, we apply two attention modules to solve the problems in different aspects, as shown in the mask attention head of Fig. 2 (b).

As shown in Fig. 3 (a), we use a variant of Squeeze-and-Excitation (SE) block [10] as the channel attention module. And different from the rough Global Average Pooling (GAP) operation in SENet, we add two convolution layers with large convolution kernel (in this paper, the kernel size is 5) before GAP to get a bigger receptive field, which helps the network to achieve better discrimination for information of the foreground and background. The following procedure is similar to SE block and the value of reduction ratio is 16. The feature outputted by the channel attention module is then concatenated with the high-level feature of the mask head. In this way, the network fuses the low-level feature full of detailed information and the high-level feature full of semantic information.

As mentioned above, the pixels on the boundaries of the instances are hard to segment correctly because the features on these regions are easily blurred by the noise or other instances. In order to effectively learn the boundary features of the instances, we introduce a supervised edge attention module based on the position attention module in DANet [6]. As shown in Fig. 3 (b), we add four convolution layers after the position attention module to get the edge attention feature on the nearly same depth as the feature that it will sum with on the mask head, which puts the two feature in an approximately equal data space and make the two feature more adaptive in the following element-wisely sum process. Then we add the up-sampled feature maps from the edge attention module and mask head together in the element-wise way to enhance features on the targeted positions, especially on the boundary of the instance.

Moreover, to guide the edge attention module to pay attention to the information of the boundary, we use a supervision mechanism by adding two convolution layers to generate the edge prediction map, and use the cross-entropy loss function with weight to calculate the loss of the edge prediction and targeted edge label. The edge label is easily obtained from the ground-truth mask label. If a pixel falls on the boundary of the object, it is considered as a positive sample and its edge label is 1 and vice versa. Since the quantities of the positive and negative samples are unbalanced, we increase the loss weights w_b of the position samples as:

$$w_b = 1/ln\,(1.2 + n/N) \tag{2}$$

Here, n denotes the number of the position samples and N denotes the total number of all the samples, including positive and negative samples.

3.3 Loss Function

We define the multi-task loss on each proposal as the sum of the losses from box head and mask head:

$$L = L_{box} + L_{seg} \tag{3}$$

where L_{box} includes three parts, that is:

$$L_{box}\,(c, B, I) = \lambda_1 L_{cls}(c, c^*) + \lambda_2 L_{reg}(B, B^*) + \lambda_3 L_{IoU}(I, I^*) \tag{4}$$

Here, L_{cls} is the focal loss for classification as in [18]. L_{reg} is the variant of GIoU loss in [24]. In origin GIoU loss, $L_{GIOU} = 1 - GIOU$ and presents anti-linear characteristic with GIoU values. In order to strengthen the hard sample learning, we use the logarithmic function to increase the losses for the samples whose GIoU values are very small.

$$L_{reg} = -\ln(\frac{GIoU + 1}{2}) \tag{5}$$

The predicted GIoU range from -1 to 1 so L_{IoU} is the l_2 loss for training.

The loss of mask head L_{seg} combines two terms:

$$L_{seg} = \lambda_4 L_{mask} + \lambda_5 L_{edge} \tag{6}$$

Here, $L_{mask}$is the loss of the mask prediction which follows the same definition as [9]. The edge label is a binary map so we define L_{edge} as the Binary Cross-Entropy (BCE) loss.

4 Experiments

We perform experiments on the large-scale object detection and instance segmentation benchmark COCO [19] and we train our models on the 115K images in COCO 2017 train set and tested on the 5K images from the validation set for ablation study. We also report our results on the 20K images in *test-dev*.

Fig. 4. Visual results of our model in COCO *test-dev*

4.1 Evaluation Metrics

We use the standard evaluation metrics in COCO, i.e., AP, AP_{50}, AP_{75} for estimating the performance of our approach on detection and segmentation. Here, AP presents average precision value over IoU thresholds from 0.5 to 0.95 with an interval of 0.05. AP_{50}, AP_{75} denote average precision at IoU thresholds of 0.5 and 0.75, respectively. Since our model deal with both object detection and instance segmentation tasks, the box AP and mask AP are all reported. For mask AP, we show the AP^m, AP^m_{50} and AP^m_{75} and for box, the AP^b, AP^b_{50} and AP^b_{75} are given.

4.2 Training Details

For fair comparison, we re-implement the baseline methods Mask R-CNN and MS R-CNN on PyTorch. The hyper-parameters in our model follows which in Mask R-CNN. And we set 16 images in a training batch (2 images per GPU) and train with Stochastic Gradient Descent (SGD) for 24 epochs with an initial learning rate of 0.02 which is decreased by a factor of 10 at the epoch 16 and 22, respectively. We train on 8 NVIDIA Tesla V100 GPUs and the weight decay and momentum are 0.0001 and 0.9. The backbones in the model are initialized with the pre-trained weights on ImageNet [5]. And we set the shorter edges of the images to be 800 and the longer edges to be 1333. The coefficient in loss function is set to $\lambda_1 = \lambda_2 = \lambda_4 = \lambda_5 = 1$, $\lambda_3 = 0.5$. As a powerful component, FPN is used in all the backbones.

Table 1. Comparison among different methods on COCO 2017 *test-dev* dataset

Method	Backbone	AP^m	$AP^m_{0.5}$	$AP^m_{0.75}$	AP^b	$AP^b_{0.5}$	$AP^b_{0.75}$
MNC [4]	ResNet-101	24.6	44.3	24.8	-	-	-
FCIS [17]	ResNet-101	29.2	49.5	-	-	-	-
FCIS+++ [17]	ResNet-101	33.6	54.5	-	-	-	-
Mask R-CNN [9]	ResNet-101-FPN	36.6	58.7	39.2	40.4	61.9	44.2
	ResNeXt-101-FPN	38.2	60.9	41.0	42.5	63.8	46.5
MS R-CNN [11]	ResNet-101-FPN	37.8	58.5	41.0	40.7	62.0	44.5
	ResNeXt-101-FPN	39.5	60.3	42.9	42.6	63.9	46.8
Mask R-CNN + ours	ResNet-101-FPN	37.7	57.6	41.1	41.7	59.6	45.8
	ResNeXt-101-FPN	39.7	60.4	43.4	44.3	62.5	48.7
MS R-CNN + ours	ResNet-101-FPN	38.6	57.8	42.4	41.7	60.4	45.4
	ResNeXt-101-FPN	40.5	60.1	44.6	44.3	62.9	48.3

4.3 Main Results

We report the performance of our proposed SEANet on *test-dev* with different backbones and frameworks in Table 1. In addition to the results of once the champions of the COCO instance segmentation challenge MNC [4] and FCIS [17] as well as its more complex version FCIS+++ [17], we also show the results of our baseline methods Mask R-CNN and MS R-CNN with different backbones for comparison. As Table 1 shows, our proposed method achieve a stably improvement on different backbones and frameworks. Especially on ResNeXt-101, our SEANet can improve the mask AP by 1.5 points for Mask R-CNN and 1 points for MS R-CNN. As for the box AP, it achieves 1.8% and 1.7% gains for Mask R-CNN and MS R-CNN respectively. Some visual results are shown in Fig. 4, which indicate the great performance of our method on instance segmentation.

4.4 Ablation Study

We conduct lots of extensive experiments to explain our components and prove the effectiveness of our approach.

The performance of the proposed SEANet. Table 2 shows that our proposed SEANet can improve both the accuracy of the instance segmentation and object detection with the gains of 1.2 AP and 1.5 AP comparing to the baseline Mask R-CNN. And from Table 2, we can see that the attention with edge supervision (the 5th row) achieves an improvement of 1.1 AP for $AP^m_{0.75}$ compared with the one without edge supervision (the 6th row). It indicates the introduction of the edge supervision in the attention module is effective for the accuracy of the instance segmentation especially for the more precise criteria $AP^m_{0.75}$. Table 2 also indicates that the improvement of the detection tasks can promote the segmentation performance, which brings a great improvement of 1.8 AP for $AP^m_{0.75}$ under the joint effect of the two parties.

Table 2. The detection and instance segmentation results on COCO 2017 validation set. The first row is the baseline Mask R-CNN framework. The component with ✓is added to the baseline. The results show the stable improvement of the proposed approach

Method	FCN box head	+ B-IoU	Attention	L_{edge}	AP^m	$AP^m_{0.5}$	$AP^m_{0.75}$	AP^b	$AP^b_{0.5}$	$AP^b_{0.75}$
Mask R-CNN (ResNet-101-FPN)					36.5	58.1	39.0	40.3	61.5	44.1
	✓				36.7	57.6	39.6	41.5	60.2	44.9
		✓			36.9	57.8	39.5	40.4	61.5	44.3
	✓	✓			37.0	57.5	39.9	41.5	60.1	45.0
		✓		✓	37.5	58.8	40.6	40.8	61.7	44.8
	✓	✓		✓	37.7	57.8	40.8	41.7	60.0	45.6

Table 3. The detection and instance segmentation results on COCO 2017 validation set. The results without ✓is the baseline method and those with ✓is our proposed approach based to the baseline. The results show the stable improvement of the proposed approach based on different backbones and different framework

Method	Backbone	SEANet	AP^m	$AP^m_{0.5}$	$AP^m_{0.75}$	AP^b	$AP^b_{0.5}$	$AP^b_{0.75}$
Mask R-CNN	ResNet-50-FPN		34.5	55.8	36.7	38.0	58.9	42.0
		✓	36.0	55.4	39.2	39.4	57.7	42.7
	ResNet-101-FPN		36.5	58.1	39.0	40.3	61.5	44.1
		✓	37.7	57.8	40.8	41.7	60.0	45.6
	ResNeXt-101-FPN		37.7	59.9	40.4	42.0	63.1	46.1
		✓	39.4	60.1	42.8	44.1	62.6	48.1
MS R-CNN	ResNet-50-FPN		36.0	55.8	38.8	38.6	59.2	42.5
		✓	36.9	55.5	40.3	39.6	58.2	42.8
	ResNet-101-FPN		37.8	58.0	41.1	40.7	61.1	45.5
		✓	38.5	57.0	42.5	41.8	59.6	45.7
	ResNeXt-101-FPN		38.8	59.5	41.7	42.1	63.1	46.0
		✓	39.8	59.3	43.7	43.7	62.3	47.6

In addition, we report the results of our proposed approach with different frameworks and backbones in Table 3, which shows that our proposed approach is universally effective and is no sensitive for the change of frameworks and backbones. In the meantime, we can see that the improvement of our method based on MS R-CNN is slightly less than the one based on Mask R-CNN. It is reasonable that since the final scores of the detected objects align with the accuracies of their classifications and regressions, which somewhat determines the qualities of the detected masks, using these scores to rank the mask results is viable. And MS R-CNN aims to find the more precise scores for the masks, in which the room for improvement is small because our box scoring is also accuracy. And the FCN box head brings more room for the edge attention module to make an impact.

The design of mask attention head. We extend the original mask head in Mask R-CNN by adding the attention modules. The results of different designs are reported in Table 4. From Table 4, we can see that using the edge supervision mechanism obtains robust performance gain compared with no applying edge

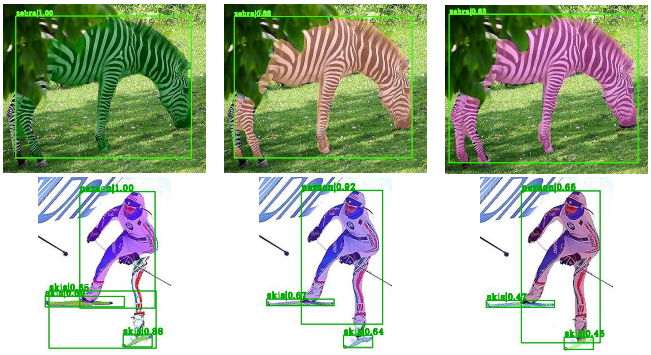

Fig. 5. Performance of the fully convolutional box head for ResNet-101-FPN. (a) The result of baseline Mask R-CNN. (b) The result of proposed SEANet when B-IoU branch is not used. (c) The result of proposed SEANet using fully convolutional box head

Table 4. Results of different designs of the attention branches. The backbone is ResNet-50-FPN

Design	AP^m	$AP^m_{0.5}$	$AP^m_{0.75}$
Mask R-CNN	34.5	55.8	36.7
(a) edge supervison (sum)	**36.0**	**55.4**	**39.2**
(b) edge supervison (concat)	35.6	55.0	38.8
(d) without edge supervison	35.1	54.6	38.4

supervision, which shows the effective of the proposed edge supervision mechanism. In the detail, the manner that element-wisely adding the feature from the edge attention module with the high-level feature on the mask head, represented by "sum", outperforms using concatenation operation which represented by "concat". It indicates that the sum operation between the feature with edge information of the object and the high-level semantic feature on the mask head can fuse the object boundary information and semantic information more directly, and specifically strengthen features on the object boundary in the pixel level.

The performance of fully convolutional box head. In order to indicate the precision with our proposed fully convolutional box head, especially the B-IoU branch, we show the visual results compared with baseline Mask R-CNN in Fig. 5. We can observe that in Mask R-CNN, in spite of giving the high classification scores, the regressions of the bounding boxes are not accurate enough and fail in locating the objects with large aspect ratios, like the skis in the Fig. 5 (a). In our proposed SEANet, even if not using the IoU scores predicted by B-IoU branch, the box regression results are evidently greater than the baseline and reasonably improve the performance of the segmentation, as shown in Fig. 5 (b). With the B-IoU branch, we multiply the IoU scores from the B-IoU branch with the classification scores to balance the precision in the aspect of regression

Table 5. The boundary prediction evaluation of the baseline Mask R-CNN and our approach. The backbone is ResNet-101-FPN

Band width	1	3	8
Mask R-CNN	7.01	29.16	41.47
Mask R-CNN + ours	**10.99**	**42.48**	**57.21**

Table 6. The comparison of runtime and parameter. The backbone is ResNet-101-FPN

Method	$runtime(ms)$	$params(M)$
Mask R-CNN	405	63.17
Cascade Mask R-CNN	507	99.09
ours (FCN box head + B-IoU)	450	58.04
ours (edge attention head)	428	71.32
ours (all)	464	64.89

and classification, which helps us to choose the most exact bounding box for each proposal. As Fig. 5 (c) shows, the borders of the detected bounding boxes press close to the boundary of each instance even if the large ratio or scale of the instance, which presents the effectiveness of the proposed fully convolutional box head.

The boundary prediction evaluation. We evaluate our proposed method on the boundary prediction by adopting a trimap approach [13]. We compute the classification accuracy within a band of varying width around boundaries of the ground-truth mask. We set three different widths of the bands and show the results of our proposed SEANet and baseline Mask R-CNN on COCO 2017 validation set in Table 5.

As Table 5 shows, our approach can significantly improve edge quality and achieve an accuracy gain of 15.74 under the band width of 8, which shows the effective of our propose method.

The comparison of runtime and parameter. Table 6 summarizes the runtimes and the model sizes of our model and other related neural network methods (Mask R-CNN and Cascade Mask R-CNN). We record runtime with $1280 * 800 * 3$ data in a batch size 8 using PyTorch 1.11.0 with a single NVIDIA Tesla V100 GPU. It can be observed that the runtime of our model is not much bigger than that of Mask R-CNN, and is smaller than that of Cascade Mask R-CNN. In terms of parameters, it is equivalent to Mask R-CNN and much smaller than Cascade Mask R-CNN.

5 Conclusions

In this paper, we focus on keeping the boundary of the mask complete in instance segmentation. We propose a fully convolutional box head with a new "B-IoU"

branch which learns association between object features and detected bounding boxes for estimating the quality of the bounding box. In the mask head, we utilize two attention modules which prove useful for feature representation. Especially, we design a novel supervised edge attention module, which use the additional supervision information to refine the edge of the mask. Experimental results show that our method consistently improves the baseline models on the COCO benchmark dataset. Since the method is benefit to both the detection and instance segmentation, we hope the method can be extended to other models dealing with instance-level object identification tasks.

Acknowledgments. This work was partially supported by the State Key Program of National Natural Science of China (No. 61836009), the National Natural Science Foundation of China (Nos. U1701267, 61871310, 61773304, 61806154, 61802295 and 61801351), the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No. B07048), the Major Research Plan of the National Natural Science Foundation of China (Nos. 91438201 and 91438103).

References

1. Arnab, A., Torr, P.H.S.: Bottom-up instance segmentation using deep higher-order crfs. In: BMVC. BMVA Press (2016) 1, 3
2. Brabandere, B.D., Neven, D., Gool, L.V.: Semantic instance segmentation with a discriminative loss function. CoRR **abs/1708.02551** (2017) 1, 3
3. Dai, J., He, K., Li, Y., Ren, S., Sun, J.: Instance-sensitive fully convolutional networks. In: ECCV (6). Lecture Notes in Computer Science, vol. 9910, pp. 534–549. Springer (2016) 4
4. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. In: CVPR. pp. 3150–3158. IEEE Computer Society (2016) 10
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. IEEE Computer Society (2009) 9
6. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: CVPR. pp. 3146–3154. Computer Vision Foundation / IEEE (2019) 2, 5, 7
7. Gao, Z., Xie, J., Wang, Q., Li, P.: Global second-order pooling convolutional networks. In: CVPR. pp. 3024–3033. Computer Vision Foundation / IEEE (2019) 5, 6
8. Hariharan, B., Arbeláez, P.A., Girshick, R.B., Malik, J.: Simultaneous detection and segmentation. In: ECCV 1
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: ICCV. pp. 2980–2988. IEEE Computer Society (2017) 1, 2, 4, 8, 10
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. pp. 7132–7141. IEEE Computer Society (2018) 5, 7
11. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: CVPR. pp. 6409–6418. Computer Vision Foundation / IEEE (2019) 4, 10
12. Jiang, B., Luo, R., Mao, J., Xiao, T., Jiang, Y.: Acquisition of localization confidence for accurate object detection. In: ECCV (14). Lecture Notes in Computer Science, vol. 11218, pp. 816–832. Springer (2018) 4
13. Kohli, P., Ladicky, L., Torr, P.H.S.: Robust higher order potentials for enforcing label consistency. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA. IEEE Computer Society (2008) 13
14. Kong, T., Sun, F., Liu, H., Jiang, Y., Shi, J.: Foveabox: Beyond anchor-based object detector. CoRR, volume = 2
15. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: ICML. pp. 282–289. Morgan Kaufmann (2001) 3
16. Li, X., Zhong, Z., Wu, J., Yang, Y., Lin, Z., Liu, H.: Expectation-maximization attention networks for semantic segmentation. CoRR **abs/1907.13426** (2019) 2
17. Li, Y., Qi, H., Dai, J., Ji, X., Wei, Y.: Fully convolutional instance-aware semantic segmentation. In: CVPR. pp. 4438–4446. IEEE Computer Society (2017) 10
18. Lin, T., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2999–3007. IEEE Computer Society (2017) 8
19. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV (5). Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer (2014) 8

20. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440. IEEE Computer Society (2015) 4
21. Novotný, D., Albanie, S., Larlus, D., Vedaldi, A.: Semi-convolutional operators for instance segmentation. In: ECCV (1). Lecture Notes in Computer Science, vol. 11205, pp. 89–105. Springer (2018) 1, 3
22. Pinheiro, P.H.O., Collobert, R., Dollár, P.: Learning to segment object candidates. In: NIPS. pp. 1990–1998 (2015) 1, 4
23. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS. pp. 91–99 (2015) 2, 4
24. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I.D., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR. pp. 658–666. Computer Vision Foundation / IEEE (2019) 8
25. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: fully convolutional one-stage object detection. CoRR **abs/1904.01355** (2019) 2, 5
26. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: ECCV (7). Lecture Notes in Computer Science, vol. 11211, pp. 3–19. Springer (2018) 5
27. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: ECCV (13). Lecture Notes in Computer Science, vol. 11217, pp. 334–349. Springer (2018) 2
28. Zhu, C., He, Y., Savvides, M.: Feature selective anchor-free module for single-shot object detection. In: CVPR. pp. 840–849. Computer Vision Foundation / IEEE (2019) 2