Joint COCO and Mapillary Workshop at ICCV 2019: COCO Panoptic Segmentation Challenge Track

Technical Report: Eyecool

Qi Wang, Hailong Zhang, Shi Ma, Yuanshuai Wang, Mei Yang, Jing Li, Feng Li
Department of Mathematics, College of Sciences, Northeastern University
NO.3-11 Wenhua Road, Heping District, Shenyang, Liaoning Province, China
wangqimath@126.com

Wuming Jiang
Beijing Eyecool Technology Co., Ltd
8th Floor, Building 1, Huihuang International Building, Shangdi 10th Street, Haidian District, Beijing
jiangwuming@eyecool.cn

Xiangde Zhang
Department of Mathematics, College of Sciences, Northeastern University
NO.3-11 Wenhua Road, Heping District, Shenyang, Liaoning Province, China
zhangxiangde@mail.neu.edu.cn

Abstract

Since the panoptic segmentation task includes both semantic and instance segmentation, most of the current panoptic segmentation networks are composed of two branches, which are processed simultaneously. The biggest problem with this structure is the lack of communication between two branches, which leads to the inadequate fusion of the two branches and the inadequate utilizing the information of the other branch. Therefore, improving the information exchange between the two branches is an important strategy to improve the accuracy of panoptic segmentation. Based on this idea, two effective branch fusion methods are proposed in this paper, namely, RPN-FCN fusion module and DAM module. Our developed method has achieved significant improvements on the COCO dataset with adding few parameters. Moreover, our method achieves 47.1% (val) / 47.1% (test-dev) on the COCO panoptic dataset, and ranks 3th on the COCO Panoptic Segmentation test-dev2019.

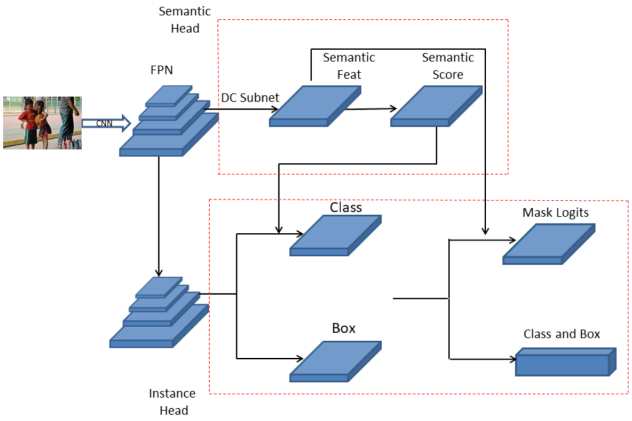

Figure 1: The illustration of overall framework. Within the two red dotted boxes there are semantic and instance branches respectively. For clarity, the final output of the semantic branch and the final fusion process of the two branches are not shown in the figure. As shown in the figure, we add two fusion approaches of two branches, where the semantic score is used to assist classification in RPN and the semantic feature is used to improve mask segmentation.

1. Overall Framework

In this section,we introduce the overall framework of our network. The overall model architecture is shown in Fig. 1.

Backbone: We adopt ResNet-152 with FPN [3] as our feature extraction network, which is commonly used in semantic and instance segmentation.

Instance Branch: For the instance branch, we adopt the Mask R-CNN [2] as our network framework, which has excellent performance on most of datasets. The main purpose of this branch is to distinguish different instances and classify them for the final panoptic fusion.

Semantic Branch: The goal of semantic branch is to classify each pixel into different categories. We use the P_2, P_3, P_4 and P_5 feature maps from FPN. Their channel numbers are 256 and their sizes are 1/4, 1/8, 1/16 and 1/32 of the original scale respectively. We apply DC subnet [6, 1] to get four new feature maps. These maps are upsampled to 1/4 of the original scale. Then we concatenate them to a new feature map. This feature map is used to obtain the final result of the semantic branch. The results of this branch will be used for panoptic fusion and the enhancement of instance branch, which will be explained in detail later.

Branch Fusion: The heuristics for merging the semantic and instance segmentation predictions follows the UPSNet [6] design. This method allows the network to classify a pixel as the unknown class instead of making a wrong prediction.

2. Fusion Module

In order to enhance the information fusion of semantic and instance branches, we propose several ingenious and effective fusion methods.

2.1. RPN-FCN Fusion Module

In the task of object detection, false and missed detections are common mistakes. These mistakes also exist in the instance branch of panoptic segmentation. Although FPN can effectively alleviate the problems on multiple scales, they still exist widely in object detection task. Especially in the RPN phase, it needs to classify several boxes for each pixel. Because of the lack of information, this task is too difficult to cause classification errors. If we classify a background box as a foreground, we still have the opportunity to correct it in the RCNN stage. If we misjudge a foreground box as a background, it will directly cause missed detection, so it is difficult to solve it just relying on instance branch. As the panoptic segmentation network has semantic branch besides instance branch, we propose a method to suppress false and missed detections by utilizing the information of semantic branch. The details are as follows.

In the RPN process, each pixel of the feature map is classified according to the IoU between the box generated around this pixel and the ground truth. For the most boxes generated by feature maps, if the box belongs to a positive case, the center of the box will be classified as foreground in semantic branch. On the contrary, the center of the box will be classified as the background in semantic branch. Similarly, the box generated at the pixel of foreground is more likely to be positive, and the box generated at the pixel of the background category mostly belongs to the negative case.

In the semantic branch we can get the segmentation map S of 133 classes, of which 80 are the background and 53 are the foreground. So we can distinguish positive and negative boxes by RPN with the help of S. S is 1/4 of the original scale. RPN uses the feature maps of FPN whose outputs are 1/4, 1/8, 1/16, 1/32 of the original map respectively. We first perform softmax on S to get the scores of 133 classes for each pixel. Then the scores of foreground are added to get a foreground (FG) score map. This score indicates the probability that the point belongs to foreground, as well as the probability that the boxes produced by the point belong to the positive case. Then, four FG maps with the same size as the FPN feature maps are obtained by downsampling the FG score map respectively.

In the RPN process, we first get the score map of the box classification, and then sigmoid is calculated on the score map to generate the probability map. The score map will have a loss with the ground truth, and the probability map is used to rank the boxes and determine whether the boxes will be processed in the next stage.

In our fusion method, we multiply the FG map with the score map to get the fusion (FS) score. Then we perform sigmoid to obtain the probability map.Compared with original score, FS score is obtained with the help of FG map. The score of foreground will be multiplied by the larger score. The score of pixel that belongs to the background will be multiplied by the smaller value. As shown in Fig. 2. The architecture is shown in Fig. 3. Therefore, our method not only restrains the missed detection of the positive box, but also reduces the risk of detecting the box of the background. This method achieves encouraging results in practical applications.

2.2. DAM Module

We first analyze the goals of semantic and instance branches. The outputs of instance branch include a bounding box regression, a classification and a segmentation mask outputs. The goal of semantic branch is to assign a semantic class label to each pixel. Among them, the segmentation of mask and semantics branch belong to the pixel-by-pixel segmentation, which requires more local information and detailed information, while the location and classification of box need more global information. Now these two branches or the four tasks share a set of feature maps, namely P_2, P_3, P_4 and P_5 obtained from FPN. This is very contradictory. In addition, there are abundant semantic information and local features in the semantic branch, which is exactly what the segmentation of mask needs. But there is little communication between the two branches, which is very inappropriate. In order to solve these two problems at the same time, we propose a module of first distinguishing and then merging

Figure 2: As shown in the figure, in the RPN-FCN fusion module, the score of the box with the center point falling in the foreground (the green box in the figure) will be multiplied by the larger value, and on the contrary, the score of the box with the center point falling in the background (the red box in the figure) will be multiplied by the smaller value. So as to restrains the problem of missed detection and false detection.

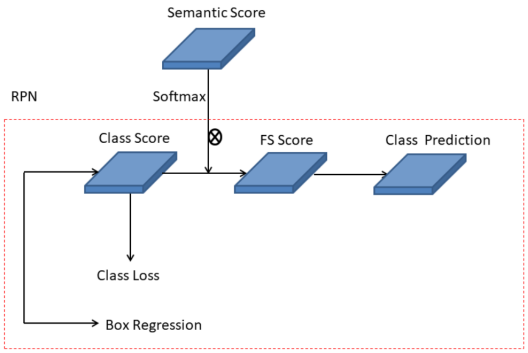

Figure 3: The illustration of RPN-FPN Fusion Module. As shown in the figure, we first perform softmax on the semantic score to obtain the foreground score map of each pixel, and we then downsample the score map to the same size as the class score. Finally, the class score is multiplied by the score map to obtain the FS score. Then we get the classification prediction by FS score. '$\otimes$' denotes multiplication of points by point correspondence.

(DAM).

We use the method of balanced instance segmentation proposed by Libra R-CNN [4]. We resize the feature maps P_2, P_3, P_4 and P_5 to the same size, for example, we resize P_2, P_3 and P_5 to the size of P_4, and then get a new feature map by calculating the average value.

$$P = \frac{1}{L}\sum_{l_{min}}^{l_{max}} Pl \tag{1}$$

We use a non-local module [5] to make the feature map contain more global information and thus better fit the instance branch. Finally, the feature map is resized to the original size of the feature map, and the specific operation is shown in the Fig. 4.

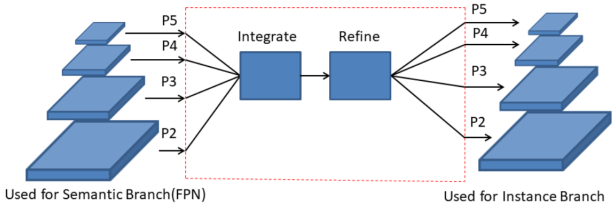

Figure 4: The illustration of Balanced Instance Segmentation.

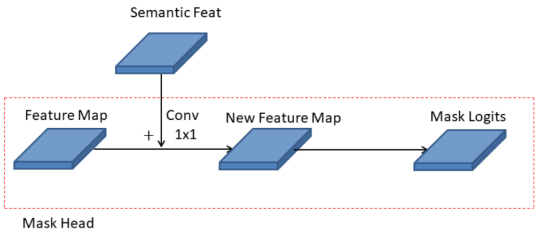

Figure 5: The illustration of DAM Module. First, the semantic feat obtained from the semantic branch has the same number of channels as the feature map from FPN by a 1x1 convolution. The semantic feat is then added to the feature map to get the new feature map. The new feature map contains a wealth of local semantic information, which is then used for mask segmentation.

After distinguishing the feature maps of semantic and instance branches, the feature maps of instance branch contain more global information, and the feature maps of semantic branch contain more semantic information. We want the instance branch to primarily handle tasks that require global information, and the semantic branch handles tasks that require local information. The goal of semantic branching is pixel-by-pixel segmentation of the whole image, which requires more local information, so no changes are needed. However, the mask segmentation of instance branch obviously does not meet our expectations, so we want to get the information needed for mask segmentation from semantic branch. On the one hand, we make full use of the information of semantic branch to achieve the communication and fusion between the two branches, thus enhancing

Models	Backbone ResNet		DAM Module	RFF Module	PQ	SQ	RQ	PQ^{Th}	SQ^{Th}	RQ^{Th}	PQ^{St}	SQ^{St}	RQ^{St}
UPSNet	50	152											
√	√				42.5	77.9	50.5	44.8	78.0	54.6	33.4	77.7	41.6
√	√		√		43.2	79.4	53.2	49.5	80.3	60.8	33.6	78.1	41.7
√		√	√		46.7	80.4	57.0	53.0	81.3	64.4	37.3	79.2	45.8
√		√	√	√	47.1	80.8	57.1	53.4	81.9	64.3	37.6	79.2	46.2

Table 1: Panoptic segmentation results on COCO.'RFF Module' means the RPN-FPN Fusion Module. We first validate the validity of the method on the ResNet-50, and finally adopt ResNet-152 as the final version.

the amount of information obtained by mask segmentation. On the other hand, we alleviate the contradiction of the instance branch and make the instance branch more suitable for its feature map. So we adopted a simple but effective fusion method—plus. Firstly, we get a feature map from the semantic branch by a 1*1 convolution, then resize it to different sizes to get a set of feature maps with the same size as the FPN feature maps. When we input the resized feature map to the mask branch, we merge it with the instance branch feature map.The architecture is shown in Fig. 5.

3. Experiments

In this section, we present the experimental results on COCO dataset. The ablation studies are carried out on COCO validation dataset as listed in Table 1. The √ in the table indicates that this module is used.

In the second row of Table 1, we adopt the DAM module. Compared with the first row, we can see that the use of the DAM module increases PQ, PQ^{Th} and PQ^{ST} by 0.7%, 4.7% and 0.2% respectively. The dramatic improvement of PQ^{Th} proves that our analysis (semantic information helps the instance branch) is correct.

Using a deeper backbone usually results in better performance. We used ResNet-152 in the third row of Table 1. Compared with the second row, we can see that the deeper backbone increases PQ, PQ^{Th} and PQ^{ST} by 3.5%, 3.5%, 3.7% respectively.

We adopt the RFF module in the fourth row of Table 1. Compared with the third row, we can see that the RFF module increase PQ, PQ^{Th} and PQ^{ST} by 0.4%, 0.4%, and 0.3% respectively. This proves that our method is effective. Because of the time limit, we only trained one epoch for the RPN-FCN fusion module, and the results may be improved more if the training process continues.

4. Conclusion

In this paper, we propose the RPN-FPN fusion module and the DAM module. Our method makes full use of the information of the semantic branch and enhance the communication between the semantic branch and the instance branch. Our method achieves 47.1% (val) / 47.1% (test-dev) on the COCO panoptic dataset, and ranks 3th on the COCO Panoptic Segmentation test-dev2019. In general, we summarize the contributions of our method as follows:

- The RPN-FPN fusion module fuses semantic information with RPN to restrain false detection and missed detection without adding additional parameters.
- The DAM module enhances instance segmentation performance by using semantic information.
- We obtain state-of-the-art performance on the COCO panoptic segmentation dataset.

References

[1] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Li Yi, Guodong Zhang, Hu Han, and Yichen Wei. Deformable convolutional networks. 2017. 2

[2] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[3] Tsung Yi Lin, Piotr Dollr, Ross Girshick, Kaiming He, and Serge Belongie. Feature pyramid networks for object detection. 2016. 1

[4] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. 2019. 3

[5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. 2017. 3

[6] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. 2019. 2