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

Technical Report: Explore Context Relation for Panoptic Segmentation

Shen Wang[1,2*] Tao Liu[1,3*] Huanyu Liu[1*] Yuchen Ma[1] Zeming Li[1] Zhicheng Wang[1]
Xinyu Zhou[1] Gang Yu[1] Erjin Zhou[1] Xiangyu Zhang[1] Jian Sun[1]

[1]Megvii Inc. [2]Peking University [3]Beijing Normal University
[1]{*liuhuanyu, mayuchen, lizeming, wangzhicheng, zxy, yugang,zej, zhangxiangyu, sunjian*}*@megvii.com*
[2]{*wangshen*}*@pku.edu.cn*, [3]{*liutao*}*@mail.bnu.edu.cn*

Abstract

This technical report presents an effective method for the panoptic segmentation. For the stuff segmentation, we propose a parallel attention module and multi-stage context branch, which take advantage of the context relations. Besides, we propose a Top-K normalization method for individual model and an adaptive scale method for the model ensemble. For the instance segmentation, we adopt a classic two-pass pipeline. Our ensemble model achieves 56.43 mIOU on the stuff segmentation validation set. By fusing the stuff segmentation and the instance segmentation result, we obtain 54.5 in PQ on the test-dev set.

1. Introduction

This technical report aims to share the details of our method for panoptic segmentation. For the stuff segmentation, we enhance the pyramid pooling module [6]. Besides, the ASPP [1] module is used to enlarge the local receptive field, and the attention module is used to increase global information. For the instance segmentation model, we adopt a two-pass pipeline. Finally, the stuff segmentation and instance segmentation are merged into the final panoptic segmentation.

In general, we summarize the contribution of our algorithm as follows:

- We use parallel attention module to enrich the context relations and enlarge the network receptive field by adding more extra res-blocks at the end of backbone.
- We design a multi-stage context branch to utilize the object information and provide more context for the stuff segmentation.
- We propose an adaptive scale strategy and a Top-K normalization method for the ensemble stage, and achieve the-state-of-art performance.

2. Methods

2.1. Stuff Segmentation

For the stuff segmentation, we mainly enrich the network attention pattern and design ensemble tricks. The network structure is shown in Figure 1. We add a parallel attention module behind the backbone and add auxiliary loss in the Res-4 block. First, we enhance the pyramid pooling module (PPM) by increasing more pooling operators. The enhanced ppm is a six-level one with bin sizes of $1, 2, 3, 6, 12, 18$ respectively. Besides, the attention feature map contains the object context[5] and the atrous spatial pyramid pooling (ASPP) [1] module increases the local receptive field of the network. This combination structure increases the local receptive field of the network and contains global context information. Moreover, we also use the Expectation-Maximization Attention (EMA)[2] module as one of our attention modules to enrich the context information.

Loss Function. We design a multi-stage context branch to make use of the object context information. In the multi-stage context branch, the object and stuff are supervised at different stages. This combined loss can promote the network to use the object context as the auxiliary information for stuff segmentation. Besides, we apply online hard ex-

*The first three authors contribute equally to this work. This work is done when Shen Wang and Tao Liu are interns at Megvii Research.

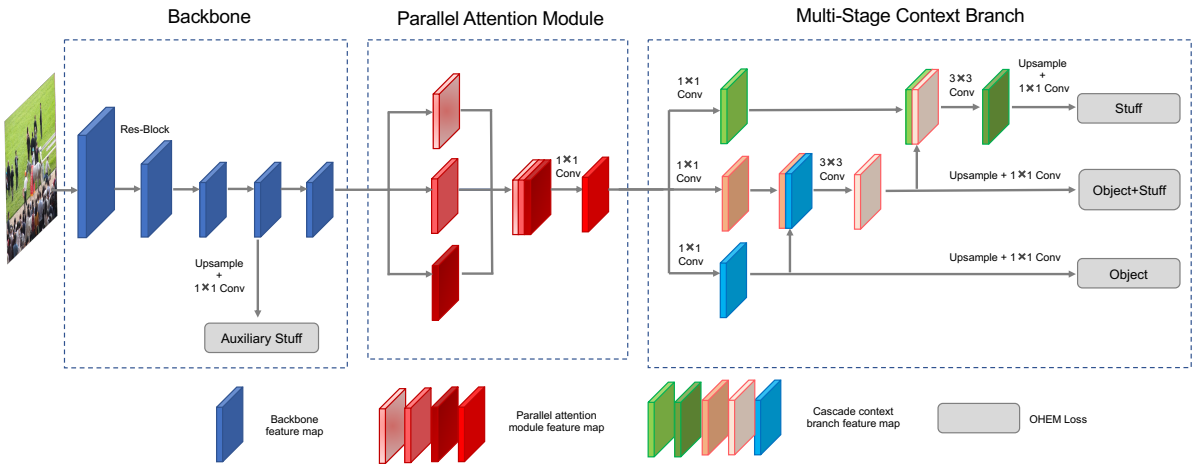

Figure 1: An overview of our proposed stuff segmentation pipeline, which is consist of three parts, backbone, parallel attention module and multi-stage context branch.

ample mining (OHEM) [4].

Model Ensemble. For the multi-scale ensemble, the same scale is usually adopted for all models in traditional methods. However, we find that adaptive scale for different models can achieve higher performance because of the model scale complement. The ablation experiments are performed in section 3.1. Besides, our statistical information indicates that the ground truth label of one pixel is generally distributed in the first k-th largest score of our model prediction, while other scores contain very little information. So we propose the **Top-K Normalization** to widen the distance between the Top-K categorys, which can increase nearly 0.1 mIOU in the final ensemble stage. The Top-K normalization operation is shown in Equation 1:

$$S_{ij,cls} = \begin{cases} S_{ij,cls} / \sum_{k=1}^{K} S_{ij,cls_k}, cls \in \{cls_1, cls_2, ..., cls_K\} \\ 0, \qquad cls \notin \{cls_1, cls_2, ..., cls_K\} \end{cases} \tag{1}$$

where $S_{ij,cls}$ is the score value in (i, j) of class cls, cls_k represents the category of the k-th largest score and K is a hyper-parameter.

2.2. Instance Segmentation

For the instance segmentation model, we adopt a two-pass pipeline. For more details, please see our instance segmentation technical report.

2.3. Panoptic Segmentation

When we get the result of the stuff segmentation and the result of the segmentation, we need to fuse to get the final result of the panoptic segmentation. It should be noted here that there is occlusion between instances [3], so we design a spatial hierarchy relation model to solve this problem. For the overlapping instances, the overlap relationship of different instances are calculated by our spatial ranking module. The instance with high ranking score is placed on top, which can effectively solve the problem of instance overlap.

3. Experiments

Dataset. We conduct all experiments on COCO panoptic segmentation dataset. This dataset contains 118K images for training, 5k images for validation, with annotations on 80 categories for the thing and 53 classes for stuff. We employ the training images for model training and test on the validation set.

Implementation Details. For a single stuff segmentation model, we use the SGD as the optimization algorithm with momentum 0.9 and weight decay 0.0001. The initial learning rate is set to 0.004. The poly learning rate policy with warm-up strategy is adopted, where the learning rate is multiplied by $\left(1 - \frac{\text{iter}}{iter_{\max}}\right)^{0.9}$. The batch size is set to 16, which means each GPU consumes two images in one iteration. We adopt some strategies for data augmentation, such as the multi-scale training, the mirror flip and random crop.

3.1. Ablation Studies

This section shows our experimental results on the COCO panoptic segmentation dataset. For the stuff segmentation, our single model and ensemble model can achieve 53.9 and 56.43 mIOU respectively on the validation set. We conduct the ablation study on the COCO validation dataset, including the parallel attention module, multi-stage context branch, adaptive scale strategy and the Top-K normalization.

PAM	MCB	Huge Backbone	mIOU
×	×	×	49.3
✓	×	×	49.9
✓	✓	×	50.4
✓	✓	✓	**53.9**

Table 1: Ablation study on COCO validation set. PAM represents Parallel Attention Module, MCB represents Multi-Stage Context Branch.

Single Model. We set the Res50 backbone with PPM module as our single model baseline, which achieves 49.3 mIOU. As shown in Table 1, parallel attention module and multi-stage context branch can improve the performance by 0.6 and 0.5 respectively. The huge backbone can further enhance the performance of our model.

Adaptive Scale Strategy. In our experiments, we first adopt the same scale $[0.8, 1.0, 1.2]$ and our ensemble model achieves 54.74 mIOU. In contrast, we set the scale to $[0.8, 1.0, 1.2]$ and $[0.8, 1.0]$ respectively for the two models and our ensemble model achieves 55.239 mIOU, which can improve performance by 0.5.

Top-K Normalization. The original ensemble of three large models achieves 55.660 mIOU. We try different K values for Top-K normalization. The experimental results show that there are best results(55.733 mIOU) when K is 4.

3.2. Final Results

Finally, the stuff segmentation and instance segmentation are merged into the final panoptic segmentation. The result of our method on the test-dev set. The experimental results are shown in Table 2.

	PQ	SQ	DQ
All	54.5	83.6	64.1
Things	64.2	86.2	74.3
Stuff	39.8	79.7	48.8

Table 2: Our final result on the test-dev set.

4. Discussion and Conclusions

This technical report presents an effective method for the panoptic segmentation task. We design the parallel attention module to get the global information in the image, and we set different scale for different models in the ensemble stage. Besides, we propose the Top-K normalization method to improve the segmentation effect. This technical report the details of our method for the panoptic segmentation.

References

[1] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[2] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. *arXiv preprint arXiv:1907.13426*, 2019. 1

[3] Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6181, 2019. 2

[4] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016. 2

[5] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018. 1

[6] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1