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

Technical Report: AntVisionPS

Weixiang Hong Qingpei Guo Wei Zhang Jingdong Chen Wei Chu
Ant Financial Services Group
{hwx229374, qingpei.gqp, ivy.zw, jingdongchen.cjd, weichu.cw}@antfin.com

Abstract

We present a panoptic segmentation (PS) system called AntVisionPS, which leverages two powerful network ends to promote both stuff and thing segmentation. The two network ends, called Enhanced UPS end (E-UPS) and Learnable Weighted Loss HTC end (LWL-HTC), are based on UPSNet [11] and HTCNet [1], respectively. In detail, we enhance UPSNet and HTCNet with several powerful components such as Libra Balanced Feature Pyramid, Location Sensitive Header and Uncertainty-weighted Loss, etc. Moreover, we propose a novel two-level end fusion strategy to further promote panoptic segmentation results from E-UPS and LWL-HTC. Without using any external data source, our AntVisionPS achieves Panoptic Quality (PQ) at 50.1% *on COCO validation set which shows the effectiveness of our system.*

1. AntVisionPS

As shown in Figure 1, our AntVisionPS is a single-backbone framework with two network ends, *i.e.*, Enhanced UPS end (E-UPS) and Learnable Weighted Loss HTC (LWL-HTC) end. Both the two ends predict semantic segmentation (SS) and instance segmentation (IS) results, which are complementary and can be merged to achieve better performance. In our experiments, E-UPS end performs well on semantic segmentation task, while LWL-HTC end excels at instance segmentation. To fully exploit the results from different ends, a novel two-level head fusion method is proposed. In training stage, we learn E-UPS and LWL-HTC separately due to memory constraint and ease of optimization. After the training is done, we inference the input image following the pipeline in Figure 1. In this section, we first present E-UPS end and LWL-HTC end separately, and then introduce the fusion strategy.

1.1. Enhanced UPS

The vanilla UPSNet [11] takes ResNet-101 [4] as the single backbone for both semantic segmentation branch and instance segmentation branch. A parameter-free panoptic head is introduced to produce panoptic segmentation result using the outputs of semantic segmentation branch and instance segmentation branch.

Our E-UPS improves existing UPSNet on three aspects. Firstly, we enhance the backbone of UPSNet by replacing ResNet-101 with SENet-154 [5]. As demonstrated in [5], SENet-154 obtains top-1 error as 16.88% on ImageNet dataset, lower than 19.87% of ResNet-101. A powerful backbone as well as its pre-trained weights on ImageNet dataset can lead to improved PQ on COCO dataset after finetuning. Secondly, we introduce Libra Balanced Feature Prymiad (BFP) [8] to strengthen the multi-level features in the FPN backbones. We utilize the non-local convolution version of Libra BFP. Thirdly, we implement Location Sensitive Header (LSH) in our network as shown in Figure 1. Our motivation is that both bounding box prediction and mask prediction are location-sensitive, while the classification prediction is supposed to be robust to location variances. Thus, it is natural to share the features of bounding box and mask prediction, while leaving classification prediction alone.

1.2. Learnable Weighted Loss HTC

HTCNet is currently the state-of-the-art IS framework. In HTCNet, a simple and straightforward SS branch is introduced for contextual information flow, which serves as a complementary for instance segmentation. However, insufficient attention was paid to the SS branch performance, making it improper for the PS task.

The LWL-HTC enhances HTCNet framework for PS task on two aspects. Firstly, we introduce uncertainty-weighted loss [6] for task-level balance between IS and SS tasks. We observe that the performance of SS is sensitive to the loss weights, while that of IS is relatively robust. Based

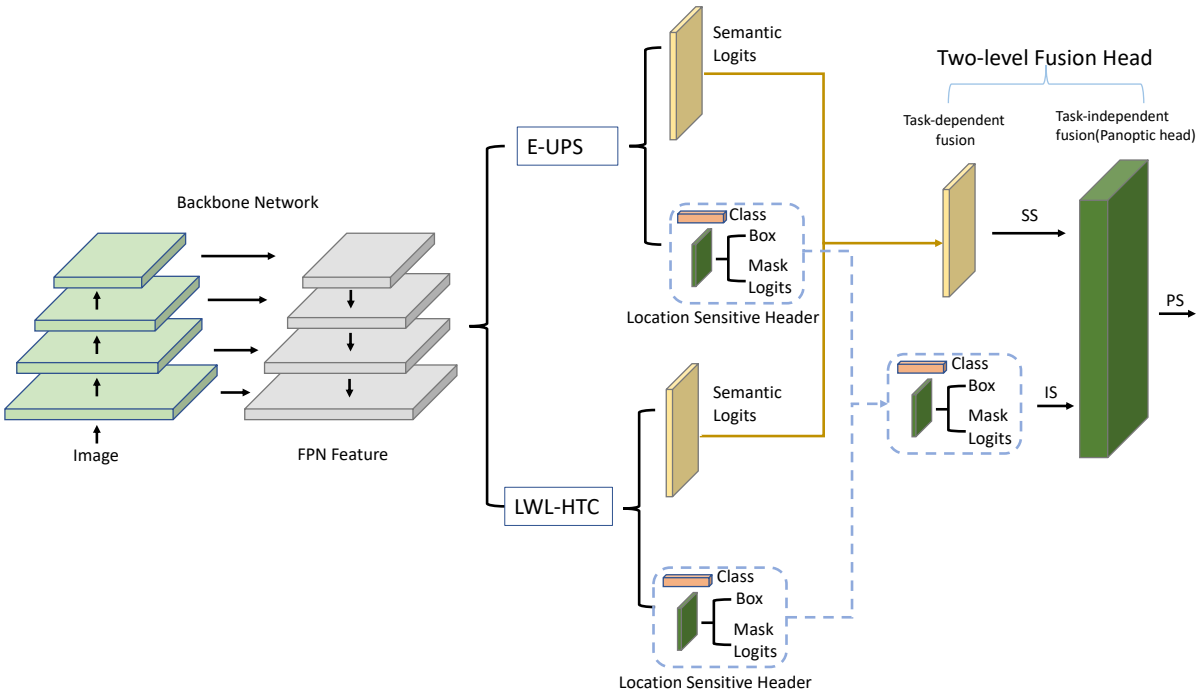

Figure 1: **System overview.**

on this observation, we appropriately adjusted loss weights to bring gains for the SS task, without hurting the IS performance. Secondly, as shown in Figure 2, we obtained a more powerful semantic segmentation head with deformable convolution for the SS task. Deformable convolution on multi-scale feature maps helps to adaptively capture context dependencies, thus brings further improvements.

1.3. Two-level Fusion Head

Since E-UPS and LWL-HTC produce semantic and instance segmentation results independently, it is difficult to merge their four outputs simultaneously. To this end, we propose a two-level fusion strategy to tackle this problem.

As shown in Figure 1, on task-dependent level, we aim to obtain better results for each sub-task. Therefore, different fusion strategies are adopted for IS and SS respectively. On the one hand, we evaluate class-wise IS performance for each head on a left out dataset in advance, fusion results are then taken from head with better performance in prior evaluation for a specific class. On the other hand, the fusion for SS results is relatively simple, we just take average of our two ends.

On task-independent level, we combine fused results from SS and IS tasks for generating final panoptic segmentation result. Here we simply adopt the heuristic combine method proposed by [7]. It is worth noting that using the panoptic head in [11] may bring potential improvements,

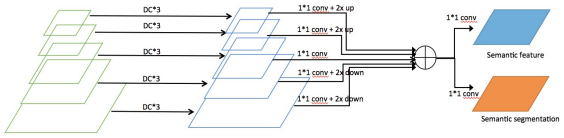

Figure 2: **Semantic Segmentation head with deformable convolution.**

and we leave it as future work.

2. Experiments

In this section, we experimentally investigate the PS performances of our AntVisionPS framework. We first measure E-UPS part and LWL-HTC part separately, then evaluate the two-level fusion method to obtain the final PQ. Additionally, we provide two ablation studies on training separate semantic segmentation branch and using larger train/test images. All results are reported based on COCO validation set.

2.1. Enhanced UPS

Table 1 shows the improvements of our modifications to the vanilla UPSNet. Libra BFP and LSH significantly boost the PQ of ResNet-101 based model from 45.0 to 46.6. Also, we observe an improvement of 0.6 after applying LSH to

SENet-154. However, due to GPU memory limitation, Libra BFP cannot work with SENet-154 together. In future, one may resort to Inplace-ABN [9] to harmonize them.

Model ensemble. We conduct model ensemble to achieve better performance of UPSNet. Specifically, we ensemble the output of four different UPSNet models based on ResNet-50, ResNet-101, ResNet-101-BFP-LSH, and SENet-154-LSH. The ensemble strategy is similar to the multi-scale ensemble method in [2], with the ensemble weights manually set as $[0.4, 1.2, 1.1, 1.3]$. After model ensemble, we obtain an increased PQ at 48.1.

2.2. Learnable Weighted Loss HTC

We investigate the effects of two main modifications in our LWL-HTC. "Uncertainty weighting" stands for the loss functions that are weighted by the homoscedastic uncertainty of SS and IS during training, while "DC head" for the utilization of deformable convolution on multi-scale feature maps in semantic segmentation branch. As shown in Table 2, "Uncertainty weighting" leads to an overall PQ gain of 1.7%, with especial boosting on PQ^{St}. Interestingly, thanks to the contextual information flow of LWL-HTC, "DC head" benefits not only SS task but also IS task, leading to the increase of PQ by 0.7%.

2.3. Two-level Fusion Head

As shown in Table 3, our fusion method outperforms the best single-model PQ by a large margin($\sim$2%), achieving 50.1% on COCO validation dataset, demonstrating the advantages of our two-level fusion method. Several illustrated samples are shown in Figure 4.

2.4. Other Attempts

Separate semantic segmentation model. Since detection model consumes memory heavily, the separation of semantic segmentation branch can enable training semantic segmentation model with large batch size that is essential to improve performance. We implement a semantic segmentation model based on HRNet [10] and DeepLabV3Plus [3]. HRNet maintains high-resolution representations which is beneficial to pixel segmentation and especially to small objects. The ASPP (Atrous Spatial Pyramid Pooling) module of DeepLabV3Plus contains several parallel atrous convolution with different rates, which can adjust filter's field-of-view in order to capture multi-scale contextual information. The decoder can successfully recover object boundaries by fusing high-level and low-level features. These characteristics are combined in our model, and the network is illustrated in Figure 3. Specifically, the last layer of HRNetV2-W48 segmentation model is replaced by ASPP module. We adopts feature maps after stage1 of HRNetV2-W48 as low level features in decoder and concatenates with output of ASPP to generate score map. The model can improve stuff segmentation results and thus improve the final panoptic segmentation results.

Table 1: **Ablation study for E-UPS.** Due to GPU memory limitation, Libar BFP cannot work with SENet-154.

	ResNet-101 with Deformable Conv.	SENet-154
vanilla	45.0	45.3
+ Libar BFP	45.7	-
+ LSH	45.8	**45.9**
+ Libra BFP & LSH	**46.8**	-
+ Model Ensemble	**48.1**	

Table 2: **Ablation study for LWL-HTC.**

	PQ^{All}	PQ^{Th}	PQ^{St}
vanilla	45.1	56.8	28.3
+ Uncertainty weighting	46.8	57.2	31.0
+ DC head	47.1	57.9	30.5
+ Uncertainty weighting & DC head	**47.5**	**58.1**	**31.6**

Table 3: **Effectiveness for Two-level Model Fusion.**

	E-UPS	LWL-HTC	Two-level Fusion
PQ^{All}	48.1	47.5	**50.1**

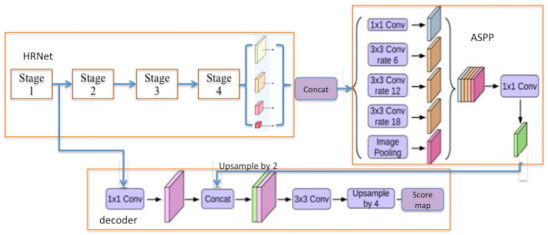

Figure 3: **Separate semantic segmentation model.**

Larger train/test image. The detection and segmentation of small objects have been the bottleneck for most detectors. A straightforward way to tackle this problem is to increase the image size during training/testing. To validate our hypothesis, we increase the train/test image size from 800 to 1024, and observe the PQ improvement of ResNet-50 from 42.4 to 43.5. Unluckily, due to GPU memory limitations, we cannot train UPSNet with larger backbones if we use train/test images at a size of 1024.

3. Conclusion

In this work, we tackle panoptic segmentation with a unified framework AntVisionPS, which has a shared backbone and two different ends. By introducing a novel fusion mechanism, AntVisionPS is capable to effectively ex-

Figure 4: **Panoptic segmentation output samples by our AntVisionPS.**

ploit the outputs of two network ends. Empirical results on COCO datasets show that our AntVisionPS achieves competitive performance compared to other methods. In the future, we would like to explore memory-efficient backbone networks and effective panoptic head.

References

[1] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 3

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[5] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[6] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018. 1

[7] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, 2019. 2

[8] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[9] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder. In-place activated batchnorm for memory-optimized training of dnns. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[10] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 3

[11] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2