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

Technical Report: Panoptic Segmentation Task

Shuchun Liu, Feiyun Zhang, Li Long, Weiyuan Shao, Jiajun Wang
The AI Lab of ELEME Inc, China
{shuchun.liu, feiyun.zhang, li.long02, weiyuan.shao, jiajun.wang }@ele.me

Abstract

This report introduces our method for the Panoptic Segmentation Task Challenge of the Joint COCO and Mapillary Workshop.Our network is built on Object detection and Stuff Segmentation to fuse them into Panoptic Segmentation. And we also take use of the method of Instance Segmentation post processing method such as CRF to further improve the performances.our final submission is an end-2-end model with two branch one instance segmentation branch and one stuff segmentation branch then to fuse them into panoptic results.Our approach achieves a performance of top6 47.0% PQ Score on COCO test-dev set.

1. Introduction

one of the main task in computer vision is to recognize all elements in an image including the class and the location of the stuff. and the high level of these elements can be dividing into two categories:things and stuff.Instance segmentation and semantic segmentation outputs are used to merge into generate panoptic segmentation predictions. While the traditional semantic segmentation output does not differentiate between different instances of things classes And the instance segmentation only get the local object information(where and what).that means that these two tasks are isolated with each other which all lack the ability to fully describe the content of the whole image.Their differences are shown in Figure 1.

To shrink the gap between these two task, the Panoptic Segmentation concept is newly published by [1] which first combined the semantic segmentation and instance segmentation tasks into one. For this new task,each pixel of an image should be assigned with only one class label and an instance id.For the things classes,the instance id is used to distinguish between different objects. And for the class label it is used to differentiate

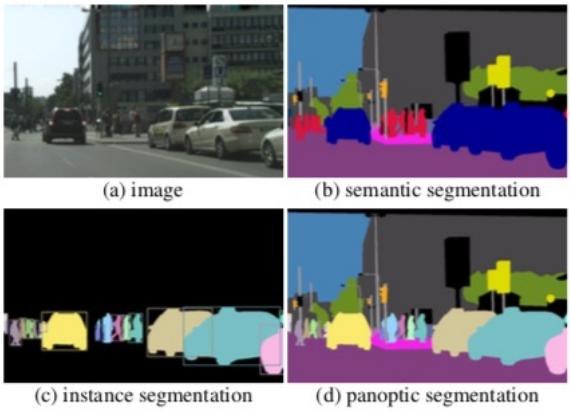

Figure 1: For the given image(a), (b)the semantic segmentation(per-pixel class labels) only show the class of each pixel,while the instance segmentation (each object mask and class label) only show the local information of a object,while (d)the panoptic segmentation can show the instance and semantic information of each pixel at the same time,which is more correspond with the human vision

between different semantic segmentation. After the task of panoptic segmentation was formally introduced,there were many new publications about this task, such as Panoptic FPN[2],AUNet[3],TASCNet[4],UPSNet[5] and OANet[6]. The task can be decomposed into three subtasks: firstly,building the network;secondly,sub task fusing,thirdly,panoptic prediction.the Panoptic FPN is mainly about the sub-task of building the network,while the AUNet & TASCNet is focusing on the sub-task fusing,and the UPSNet[5] is dealing with the panoptic prediction.

In our report, we proposed a single end-to-end Newark that makes use of both instance segmentation and semantic segmentation predictions, using a shared feature extractor. These predictions are combined to form panoptic segmentation outputs. Our main contribution is apply end-to-

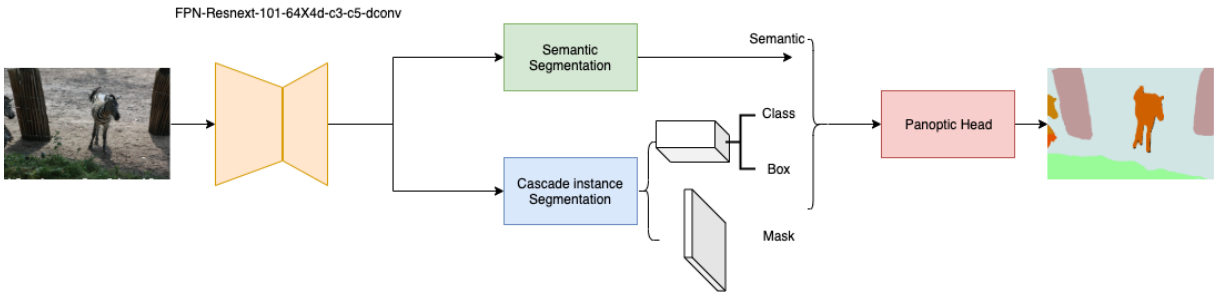

Figure 2: The Structure of Our Method

end method to make semantic segmentation and instance segmentation predictions to final panoptic segmentation results.

Method	PQ	SQ	RQ
Ours	0.470	0.815	0.563

Table 1: the results on COCO test-dev dataset

2. Method

We proposed a method that combines the semantic segmentation and instance segmentation outputs.this method consists of four main sections:Firstly,a Convolution Neural Network with feature pyramid networks; Secondly,semantic segmentation branch after the first backbone;Thirdly,the object detection branch after the backbone,Fourthly, the fusing panoptic head to mix together the semantic segmentation and object detection information into the panoptic segmentation output.

2.1. Network architecture

Our method consists of a shared convolutional feature extraction backbone and multiple heads on top of it.Each head is a sub-network which leverages the features from the backbone and serves as specific design purpose that is explained in detail below.The overall model architecture is shown in Fig2.

Backbone:The base of the networks is a Resnext-101-64X4d[7]feature extractor,which is shared by the semantic segmentation and instance segmentation branch.To enhance the ability of the model we also add three deformable layers to the backbone of FPN[8] P3,P4 and P5.

Instance Segmentation Head:In the instance segmentation branch we use the popular Cascade Mask R-CNN[9] to detect the objects. There are three stages of the Cascade Mask R-CNN,we choose the middle stage as the instance segmentation flow.

Semantic Segmentation Head: the semantic segmentation branch is constructed based on the output of the FPN[8].We use P2,P3,P4 and P5 feature maps of FPN[8] which contain 256 channels and are 1/4,1/8,1/16 and 1/32 of the original scale respectively. These feature maps first go through the same deformable convolution network independently and are subsequently up-sampled to the 1/4 scale. we then concatenate them and apply 1x1 convolutions with Softmax to predict the semantic class.

Panoptic Head:And in the final Panoptic head we use the UPSNet[5] method. It's a connection of the state-of-art detection and panoptic fusing method.And after inference we also use the Dense-CRF[10] method to fix the imperfect edge of the instance with the stuff which can also improve the performance of the model.

3. Implementation

During the training phase, the network is trained end-to-end,using SGD optimizer with the momentum of 0.9,the initial learning rate is 0.01, and the learning rate is decreased twice with a factor of 2. and trained for 20 epoch on the COCO database. The network is initialized using weights pre-trained on the ImageNet dataset, and its is always trained on 16 Tesla V100 GPUs.All presented results are from a single model.

4. Results

The results on the datasets of COCO have been submitted to the COCO and Mapillary joint Recognition Challenge 2019. At the time of submitting, the results on the challenge test sets have not yet been announced. The challenge use the Panoptic Quality(PQ)[1] to evaluate the result.This metric is designed to assess both the segmentation and recognition quality of the different methods which is composed of two targets: the Segmentation Quality(SQ) and the Recognition Quality(RQ). Our method are shown in Table.

5. Conclusion

We presented a method that is able to combine the Cascade Mask-RCNN with the segmentation predictions to produce the panoptic results. Although the performance of this method is worse than some others in the leaderboards,

but we finished the task. In the future, We want to further explore the potential benefits of joint learning for panoptic segmentation.

References

[1] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 1, 2

[2] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6399–6408, 2019. 1

[3] Yanwei Li, Xinze Chen, Zheng Zhu, Lingxi Xie, Guan Huang, Dalong Du, and Xingang Wang. Attention-guided unified network for panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7026–7035, 2019. 1

[4] Jie Li, Allan Raventos, Arjun Bhargava, Takaaki Tagawa, and Adrien Gaidon. Learning to fuse things and stuff. *arXiv preprint arXiv:1812.01192*, 2018. 1

[5] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8818–8826, 2019. 1, 2

[6] Jiahui Zhang, Dawei Sun, Zixin Luo, Anbang Yao, Lei Zhou, Tianwei Shen, Yurong Chen, Long Quan, and Hongen Liao. Learning two-view correspondences and geometry using order-aware network. *International Conference on Computer Vision (ICCV)*, 2019. 1

[7] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[9] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *arXiv preprint arXiv:1906.09756*, 2019. 2

[10] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011. 2