Supplementary Material: Single-shot Path Integrated Panoptic Segmentation

Sukjun Hwang[1] Seoung Wug Oh[2] Seon Joo Kim[1]

[1]Yonsei University [2]Adobe Research

1. Additional Results on Test Sets

We provide our scores on test sets that had to be omitted in the main paper due to the space limit. The performance of our model is measured under the same setting used for validation sets. To ease the understanding of our model's performance, we reference the tables from Panoptic-DeepLab [1] (Table 3, 8, 9). For fairness, we list the scores from the models that are pretrained only on ImageNet.

2. Network Structure Details

We find that applying the group normalization [9] until the last layers of Panoptic-Feature generator harms the training stabilization at the early stage. Therefore, we do not apply group normalization for the last layers after the absolute positional encoding.

3. Training Details

3.1. Ground Truth Assignment

The assignment policy of ground-truth to Filter Sampling Module is fairly similar to the one of FCOS [8]. Let $M_i \in \{0,1\}^{H \times W}$ be the ground-truth mask of an instance i, where H and W are the actual size of the input image. The corresponding position of the feature (x', y') to the input (x, y) can be obtained as $(\lfloor \frac{s_l}{2} \rfloor + x' * s_l, \lfloor \frac{s_l}{2} \rfloor + y' * s_l)$, where s_l denotes the stride of the feature at level l. We assign ground-truth of an instance i to the locations (l, x', y') where corresponding $M_i(x, y)$ is $true$. If multiple instances correspond to a same location, the instance of least mask area becomes the target. With this policy, SPINet can have multiple feature locations designated to a same instance, thus a collection of substantial training samples is possible.

We assign ground truth to each level by $\sqrt{h \times w}$ of the bounding box of an instance: $(0, 48)$, $(24, 96)$, $(48, 192)$, $(96, \infty)$ for Cityscapes, and $(0, 96)$, $(48, 192)$, $(96, 384)$, $(192, 768)$, $(384, \infty)$ for COCO.

Method	Backbone	PQ (%)
COCO *test-dev* set		
TASCNet [5]	ResNet-50	40.7
Pan-FPN [4]	ResNet-101	40.9
AdaptIS† [7]	ResNeXt-101	42.8
DeeperLab [11]	Xception-71	34.3
SSAP‡ [3]	ResNet-101	36.9
Pan-DL [1]	MobileNet-V3	29.8
Pan-DL	ResNet-50	35.2
Pan-DL	Xception-71	39.6
SPINet	ResNet-50	42.6
Cityscapes *test* set		
SSAP‡ [3]	ResNet-101	58.9
Pan-DL [1]	MobileNet-V3	54.1
Pan-DL	ResNet-50	58.0
Pan-DL	Xception-71	60.7
SPINet	ResNet-50	60.2

Table 1. Results on COCO *test-dev* set and Cityscapes *test* set. †: adding left-right flipped inputs. ‡: adding left-right flipped inputs and multi-scale inputs.

3.2. Data augmentation

For COCO, image scale variation of $[0.9, 1.0]$ is applied with horizontal flipping.

For Cityscapes, following steps of are applied to the input images: color augmentation, scale variation of $[0.5, 2.0]$, and crop by the absolute size of 512×1024. The crop size is adjusted to fit the batch size of 4 per GPU with 11GB of memory, which is a practical spec for general users.

4. Inference Details

The threshold of the overlap between instances is set to 0.5. The stuff segments that have smaller region than a threshold are left as *unknown*, and the thresholds for COCO and Cityscapes are 4096 and 2048 respectively.

5. Visualized Outputs on COCO

As shown in Fig. 1 we provide results on COCO [6] dataset, which had to be omitted in the main paper due to the space limit.

6. Visual Comparison to Panoptic-DeepLab

As Panoptic-DeepLab [1] is considered as a strong baseline for the panoptic segmentation task, we provide visualized outputs from our model and Panoptic-DeepLab as shown in Figure 2. The codes and weights we used for Panoptic-DeepLab is from `Detectron2` [10], and the backbone it used has the same modification to ours, described in the main paper, Sec 5.2.

References

[1] Bowen Cheng, Maxwell D. Collins, Yukun Zhu, Ting Liu, Thomas S. Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, 2020. 1, 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 4

[3] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In *ICCV*, 2019. 1

[4] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollar. Panoptic feature pyramid networks. In *CVPR*, 2019. 1

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

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 3

[7] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7355–7363, 2019. 1

[8] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, 2019. 1

[9] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 1

[10] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 2

[11] Tien-Ju Yang, Maxwell D. Collins, Yukun Zhu, Jyh-Jing Hwang, Ting Liu, Xiao Zhang, Vivienne Sze, George Papandreou, and Liang-Chieh Chen. Deeperlab: Single-shot image parser. *CoRR*, abs/1902.05093, 2019. 1

Figure 1. Examples of visualized outputs (Best viewed on a high-resolution display with zoom-in). The images are from COCO [6] dataset.

Figure 2. Examples of visualized outputs (Best viewed on a high-resolution display with zoom-in). The images are from Cityscapes [2] dataset.