Supplementary Material:
PRN: Panoptic Refinement Network

Bo Sun[1] Jason Kuen[1] Zhe Lin[1] Philippos Mordohai[2] Simon Chen[1]
[1]Adobe Inc. [2]Stevens Institute of Technology
{bosu, kuen, zlin, sichen}@adobe.com philippos.mordohai@stevens.edu

Abstract

In this document, we present implementation details of MS-PanopticFPN, which is used as a base panoptic segmentation network in our experiments (Section 1); additional qualitative results on the Cityscapes and COCO datasets (Sections 2 and 3); and more details on our ablation studies, which have been summarized in the main paper (Section 4).

1. Details of MS-PanopticFPN

As shown in Fig. 1, MS-PanopticFPN is inspired by PanopticFPN [3] and is composed of a backbone network combined with detection, instance segmentation and semantic segmentation branches. The input is an RGB image and the output is predicted per-pixel semantic and instance labels. The detection module is based on ATSS [9], modified to include a hierarchical classification head, with decoupled objectness and classification prediction heads. The detection loss consists of three parts: centerness loss, bounding box regression loss and focal loss [5] for classification. It also includes instance and semantic segmentation modules that share parameters with the detection module. The instance and semantic segmentation branches share the same FPN backbone features as the detection branch.

We use the semantic segmentation branch from Real-time Panoptic [2] for semantic segmentation. Multi-scale semantic features from the detection branch are fed to the stuff segmentation branch to predict per-pixel semantic labels for each image. Features from the classification branch are upsampled to an intermediate size of (H/4, W/4) and concatenated into a global feature F. Semantic labels are then predicted from F through a single convolutional layer. For the batch normalization layer in the stuff segmentation branch, we use the running statistics of the detection branch. In other words, we use the same mean and variance for both detection and stuff segmentation. We use dice loss [7] and focal loss [5] for the semantic segmentation branch.

For instance segmentation, we use the instance segmentation branch from CenterMask [4], which shares the same backbone with the detection branch. We feed the features from the Feature Pyramid Network (FPN) using RoI Align [1] to the instance segmentation branch and then predict per-object masks. After object proposals are predicted by the detection branch, we use RoI align to crop the features from different levels of the feature maps in FPN. We use SAG-Mask from CenterMask [4] for the instance segmentation branch. Once features inside the predicted RoI are extracted by RoI Align at 14×14 resolution, they are fed into four convolutional layers and spatial attention module (SAM) sequentially. Then, a 2×2 de-convolution upsamples the feature map to 28×28. Lastly, a 1×1 convolutional layer is applied for predicting instance masks. We use focal loss to train the instance segmentation branch.

2. Qualitative Results on Cityscapes

Figure 2 shows additional qualitative results of Real-time Panoptic [2], SegFix [8] and PRN on Cityscapes. Notice the limitations of SegFix compared to PRN in these examples.

3. Additional Qualitative Results on COCO

Figures 3 and 4 show additional qualitative results on COCO dataset [6].

4. Ablation Studies

In this section, we present Table 1, which was omitted from the main paper, and discuss the ablation studies in more detail.

The COCO validation set was used in these ablation studies to evaluate the effectiveness of each component in our refinement network. In order to merge the predicted center and offset map into the instance mask, we need the foreground mask to filter out the background pixels. We can obtain the foreground mask from the semantic segmentation branch or foreground mask branch. RPN improves the PQ of MS-PanopticFPN by 0.8% using the foreground mask

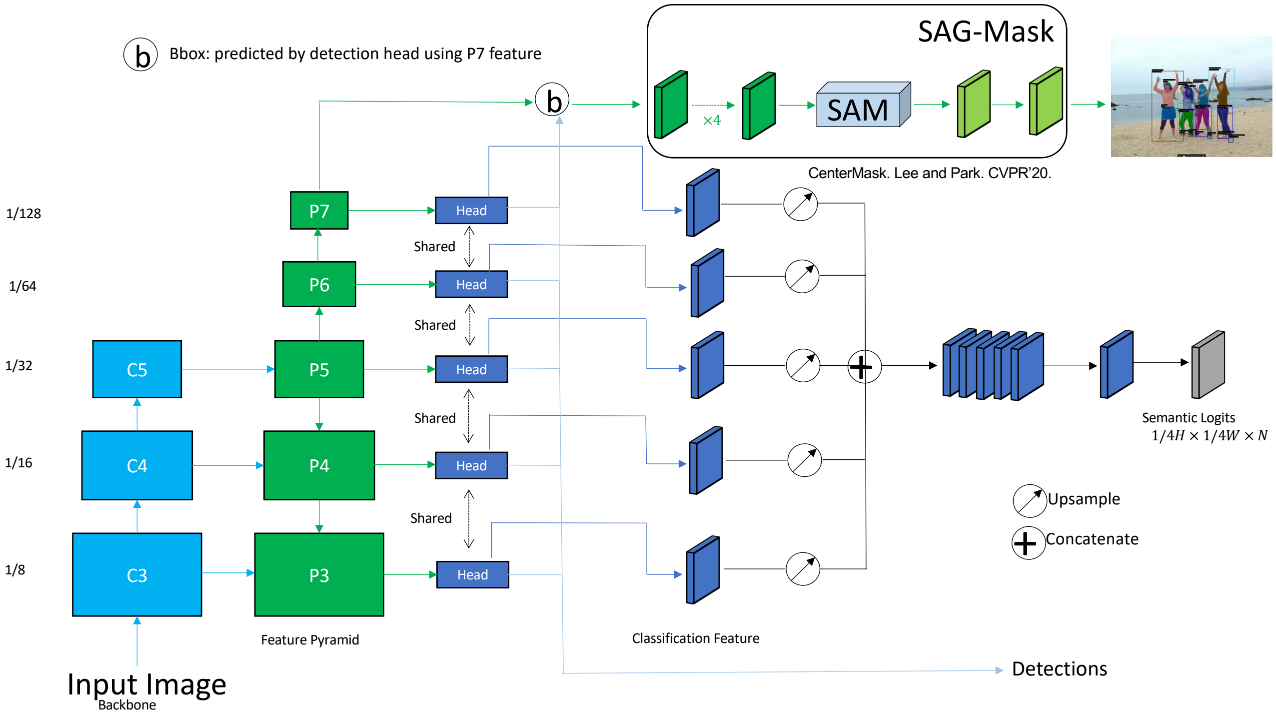

Figure 1. Overview of the architecture of MS-PanopticFPN.

Method	Foreground	ForegroundSem	CoordConvDec	CoordConvEnc	PredBbox	PQ	PQ^{Th}	PQ^{St}
MS-PFPN	-	-	-	-	-	40.6	46.6	31.6
PRN		✓				41.4	47.3	32.6
PRN	✓					41.9	47.9	33.1
PRN	✓		✓			42.2	48.1	33.2
PRN	✓			✓		42.0	47.8	33.0
PRN	✓		✓	✓		42.5	48.5	33.5
PRN	✓				✓	43.1	48.9	33.2
PRN	✓		✓	✓	✓	**44.4**	**50.9**	**34.4**

Table 1. Ablation study for PRN on the COCO validation set. Foreground means using the foreground mask from foreground mask branch. ForegroundSem means using the foreground mask from the semantic segmentation branch. CoordConvDec means applying CoordConv in the decoder layers. CoordConvEnc means applying CoordConv in the encoder layers. PredBbox means using predicted bounding box at each pixel (in addition to center and center offset maps) in the postprocessing to group instance pixels.

from the semantic segmentation branch, and by 1.3% using the foreground mask from our foreground mask branch. This justifies the inclusion of the foreground mask branch.

We also applied CoordConv at different parts of the encoder-decoder: (1) only in encoder layers, (2) only in decoder layers, (3) in both encoder and decoder layers. The PQ of our refinement network can be improved by an additional 1.4% and 1.6% if we use CoordConv in the decoder and encoder layers respectively. CoordConv works better in the decoder layers than encoder layers. We can improve the PQ by 1.9% by applying CoordConv in both the encoder and decoder layers.

We can further improve the PQ by 2.5% when we use predicted bounding boxes at each pixel when merging the center and offset maps. PQ is improved by 3.8% when we apply CoordConv in both encoder and decoder layers and use predicted bounding boxes in postprocessing.

References

[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[2] Rui Hou, Jie Li, Arjun Bhargava, Allan Raventos, Vitor Guizilini, Chao Fang, Jerome Lynch, and Adrien Gaidon. Real-time panoptic segmentation from dense detections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8523–8532, 2020.

[3] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of*

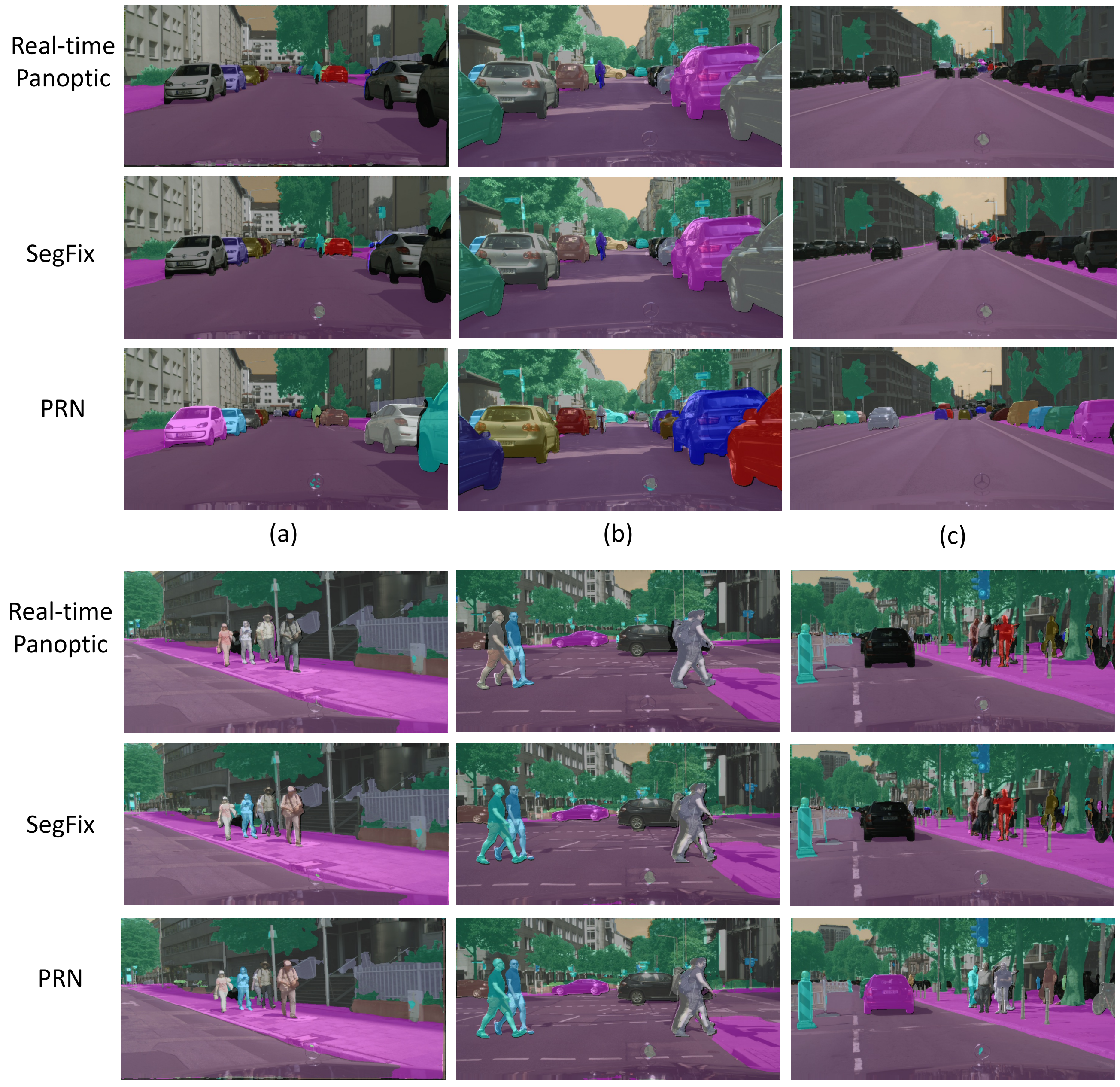

Figure 2. Qualitative results on Cityscapes dataset. Note that the color of the instance masks represents the index of each instance, and not its class label. In (a), (b), (c) and (f), PRN is able to restore the missing car, bicycle and person starting from the results of Real-time Panoptic. SegFix fails to detect the missing objects. In (d) and (e), PRN splits the person segmentation masks which are mixed by Real-time Panoptic. SegFix cannot make these corrections.

the IEEE Conference on Computer Vision and Pattern Recognition, pages 6399–6408, 2019.

[4] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2020.

[5] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[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 *European conference on computer vision*, pages 740–755.

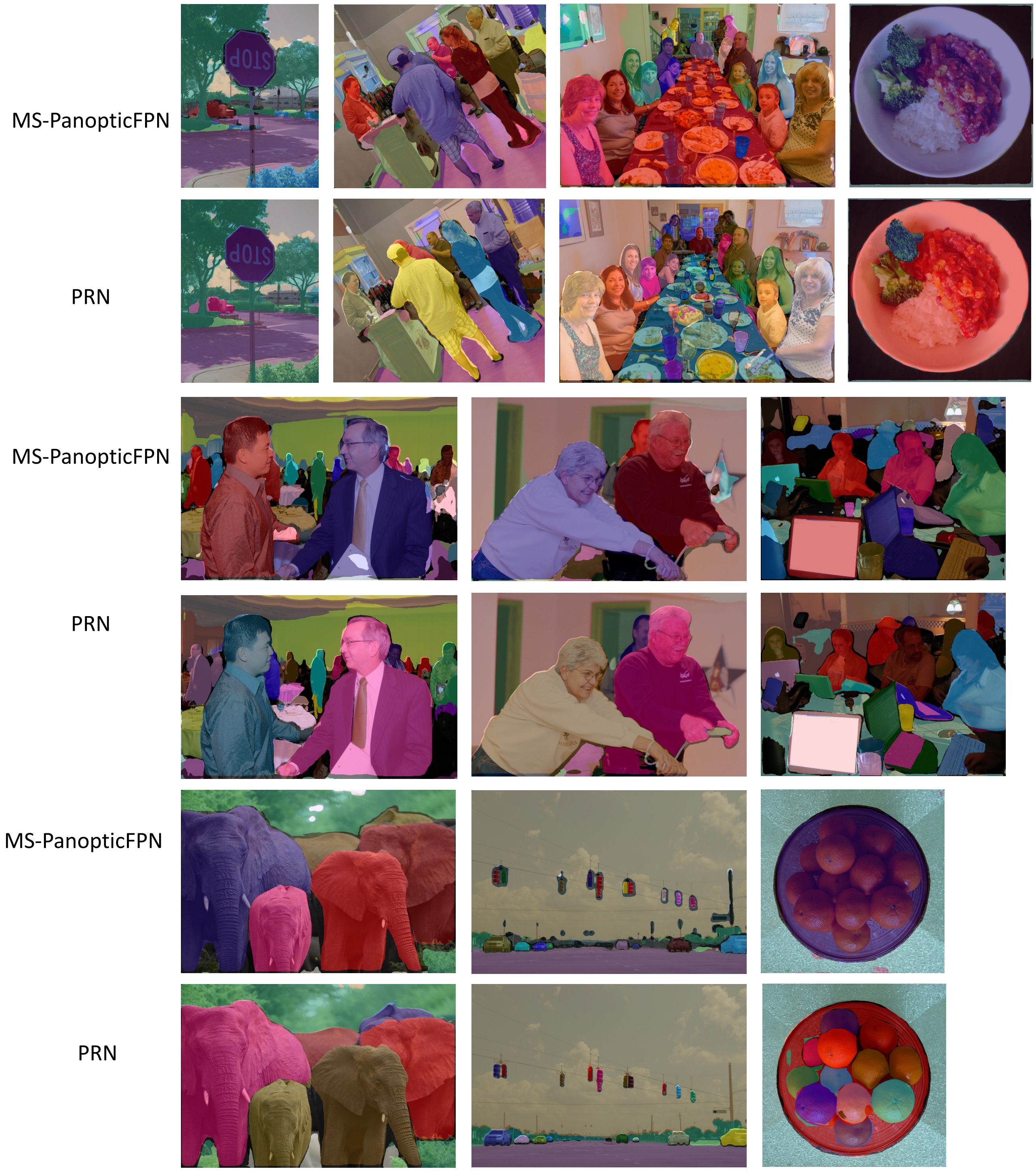

Figure 3. Additional qualitative results on COCO dataset. Note that the color of an instance mask represents the index, not the class label, of the instance.

Springer, 2014.

[7] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[8] Yuhui Yuan, Jingyi Xie, Xilin Chen, and Jingdong Wang. Seg-

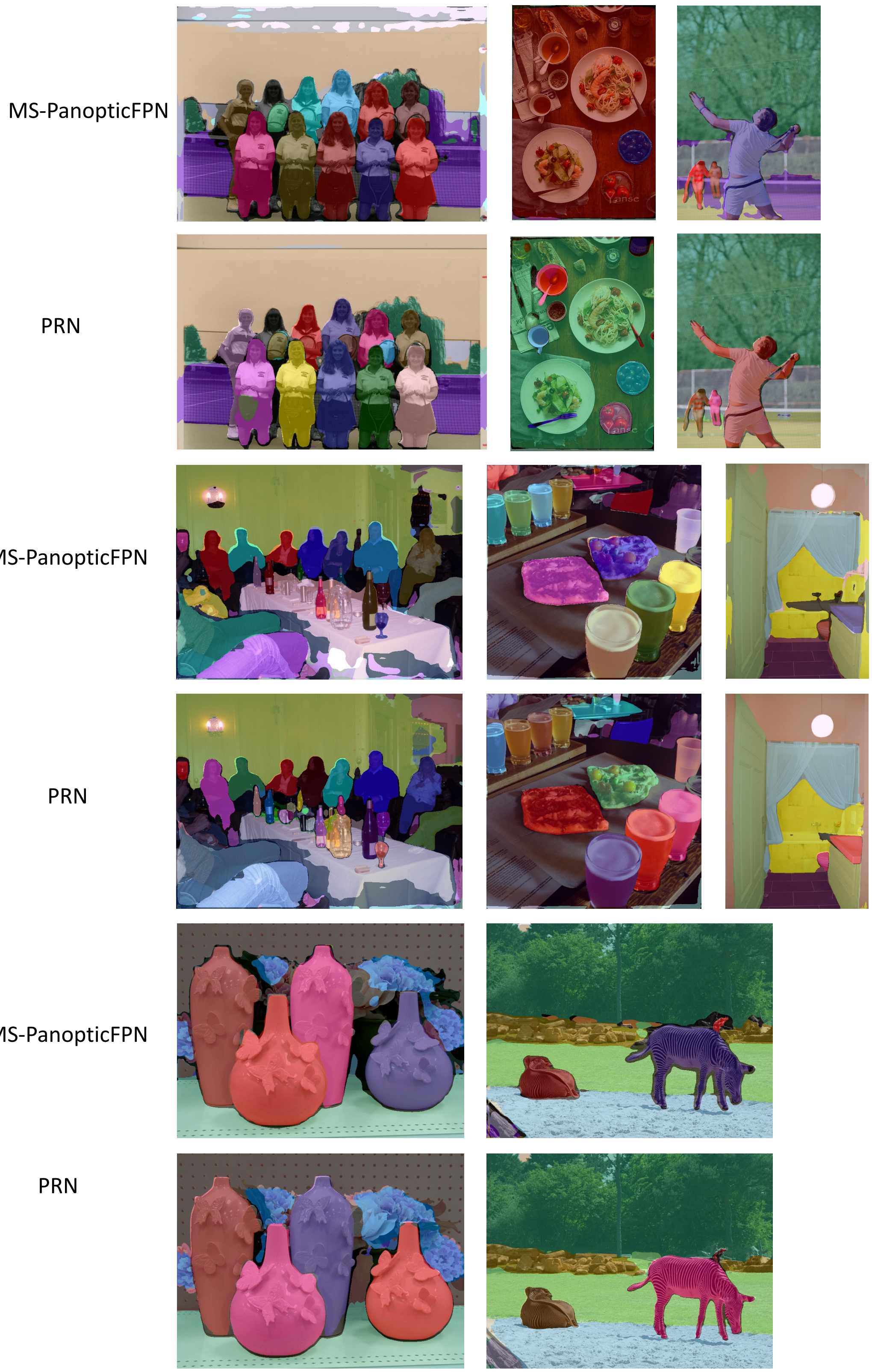

Figure 4. Additional qualitative results on COCO dataset. Note that the color of an instance mask represents the index, not the class label, of the instance.

Fix: Model-agnostic boundary refinement for segmentation. In *Eur. Conf. Comput. Vis.*, pages 489–506, 2020.

[9] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020.