Supplementary Material:
PRN: Panoptic Refinement Network

Bo Sun\textsuperscript{1} Jason Kuen\textsuperscript{1} Zhe Lin\textsuperscript{1} Philippos Mordohai\textsuperscript{2} Simon Chen\textsuperscript{1}
\textsuperscript{1}Adobe Inc. \{bosu, kuen, zlin, sichen\}@adobe.com
\textsuperscript{2}Stevens Institute of Technology 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 \cite{2} 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 \cite{3}, 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 \cite{4} 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 \cite{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 \cite{5} and focal loss \cite{4} for the semantic segmentation branch.

2. Qualitative Results on Cityscapes

Figure 2 shows additional qualitative results of Real-time Panoptic \cite{2}, SegFix \cite{6} 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 \cite{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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Foreground</th>
<th>ForegroundSem</th>
<th>CoordConvDec</th>
<th>CoordConvEnc</th>
<th>PredBbox</th>
<th>PQ</th>
<th>PQ_{th}</th>
<th>PQ_{st}</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-PFPN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>40.6</td>
<td>46.6</td>
<td>31.6</td>
</tr>
<tr>
<td>PRN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>41.4</td>
<td>47.3</td>
<td>32.6</td>
</tr>
<tr>
<td>PRN</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>41.9</td>
<td>47.9</td>
<td>33.1</td>
</tr>
<tr>
<td>PRN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>42.2</td>
<td>48.1</td>
<td>33.2</td>
</tr>
<tr>
<td>PRN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>42.0</td>
<td>47.8</td>
<td>33.0</td>
</tr>
<tr>
<td>PRN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>42.5</td>
<td>48.5</td>
<td>33.5</td>
</tr>
<tr>
<td>PRN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>43.1</td>
<td>48.9</td>
<td>33.2</td>
</tr>
<tr>
<td>PRN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>44.4</strong></td>
<td><strong>50.9</strong></td>
<td><strong>34.4</strong></td>
</tr>
</tbody>
</table>

References

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.


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.


[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.