PEM: Prototype-based Efficient MaskFormer for Image Segmentation

Supplementary Material

6. Additional Implementation Details

In the subsequent section, we offer supplementary information about the dataset utilized in our experiment, indicating some additional training parameters used for each respective dataset.

Cityscapes. Cityscapes is a dataset containing high-resolution images (1024×2048 pixels) that depict urban street views from an egocentric perspective. The dataset comprises 2975 training images, 500 validation images, and 1525 testing images, encompassing 19 distinct classes.

For both segmentation tasks, we use a fixed crop size of 512×1024 during training. During inference, the full-resolution image is used.

ADE20K. The ADE20K dataset contains 20,000 images for training and 2,000 images for validation, captured at diverse locations and featuring a wide range of objects. The images vary in size.

Following the methodology proposed in [5], a unique crop size is utilized for each segmentation task during the training process. Semantic segmentation uses a fixed crop size of 512×512, while panoptic segmentation uses a fixed crop size of 640×640. During inference, the shorter side of the image is resized to fit the corresponding crop size.

Metrics. The metric adopted for semantic segmentation, mean Intersection over Union, is defined as:

$$\text{mIoU} = \frac{1}{K} \sum_{i=1}^{K} \left| \frac{\mathcal{Y} \cap Y}{\mathcal{Y} \cup Y} \right|, \tag{11}$$

where K is the number of classes, the numerator is the intersection between the predicted mask $\mathcal{Y}$ and the ground truth Y, while the denominator is their union.

Considering panoptic segmentation, Panoptic Quality takes into account both the quality of object segmentation and the correctness of assigning semantic labels to the segmented objects. Formally, it is defined as follows:

$$\text{PQ} = \frac{\sum_{i \in TP} \text{IoU}}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}. \tag{12}$$

In practice, it can be seen as the product of Segmentation Quality (SQ) and Recognition Quality (RQ):

$$\text{PQ} = \underbrace{\frac{\sum_{i \in TP} \text{IoU}}{|TP|}}_{\text{SQ}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{RQ}}. \tag{13}$$

N	PQ	PQ_{th}	PQ_{st}	FLOPs
50	58.7	48.8	65.9	225G
100	**61.1**	**54.3**	66.1	237G
200	61.0	53.7	**66.4**	275G

Table 7. Ablation on number of queries on Cityscapes.

SQ measures the similarity between correctly predicted segments and their corresponding ground truths while PQ measures the overall ability of the model in identifying and classifying objects or segments.

Baselines. In Tab. 4, the task-specific architectures have been retrained from scratch given that these models were not tested on the ADE20K dataset. To ensure fairness in comparison with our model, we performed hyper-parameter tuning for the different architectures. Nonetheless, our comprehensive evaluation has revealed that our pipeline demonstrates suitability across diverse approaches, ultimately yielding the highest results. Specifically, we trained these models for 160,000 iterations using AdamW [23] optimizer, Cosine learning rate scheduler [22], initial learning rate set to 0.0004, and weight decay set to 0.05. The crop size remained consistent with our experiment, at 512×512.

7. Additional Ablation Study

Number of queries. The results for Cityscapes varying number of queries are reported in the Tab. 7. Using 50 queries leads to a performance drop while increasing them to 200 is not beneficial while causing a substantial increase in complexity.

8. Qualitative Results

We showcase qualitative results of PEM on the Cityscapes and ADE20K datasets, highlighting its performance both in semantic and in panoptic segmentation. Our evaluations involve comparisons with resource-intensive architectures like Mask2Former [5] and lightweight alternatives such as YOSO [18]. PEM exhibits comparable performance to Mask2Former while demonstrating superiority over YOSO. Specifically, our model excels in distinguishing different instances in the panoptic setting and displays a lower number of false positives.

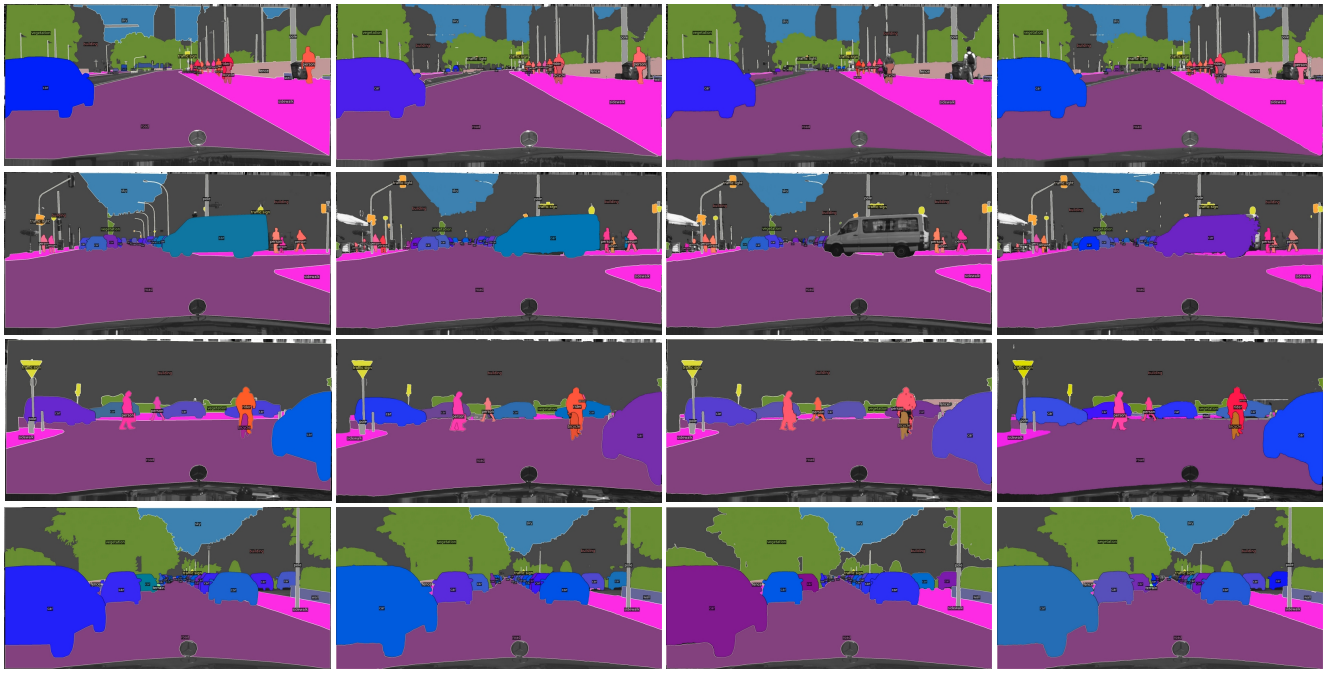


Figure 6. **Qualitative results** of PEM *v.s.* Mask2Former [5] and YOSO [18] on *panoptic segmentation* on Cityscapes.

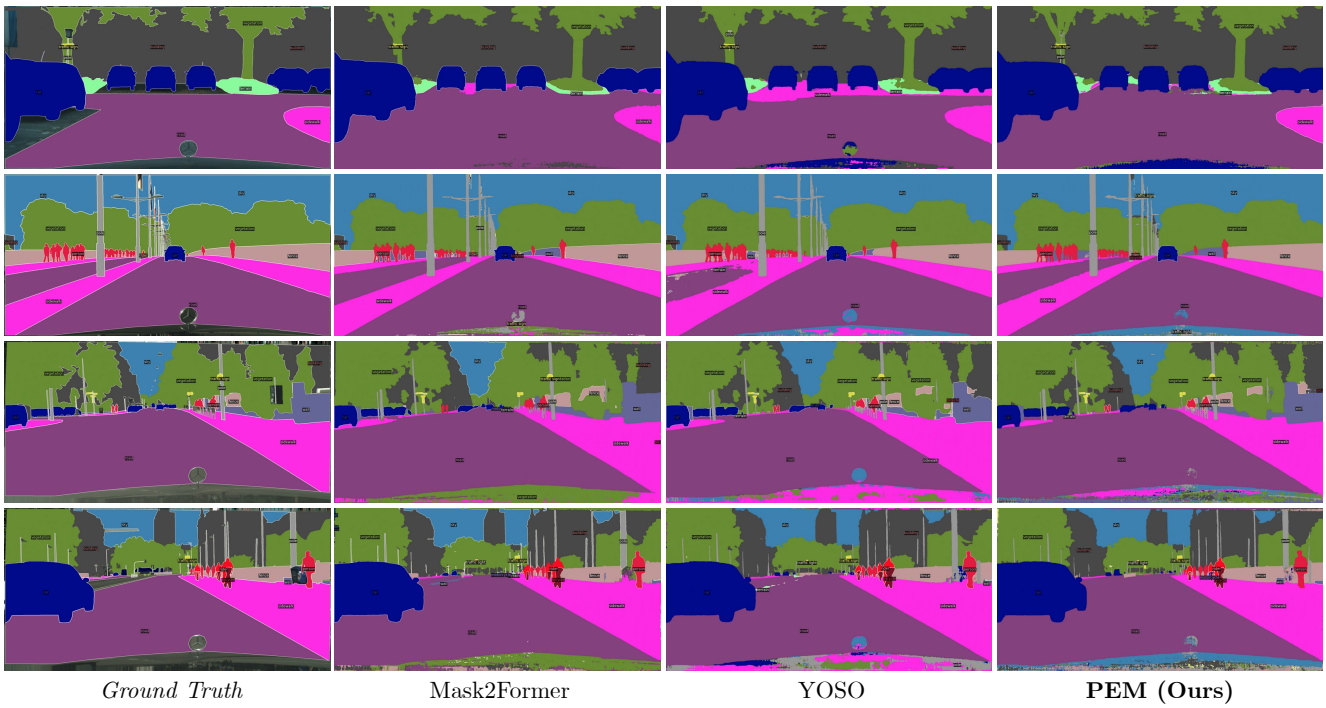
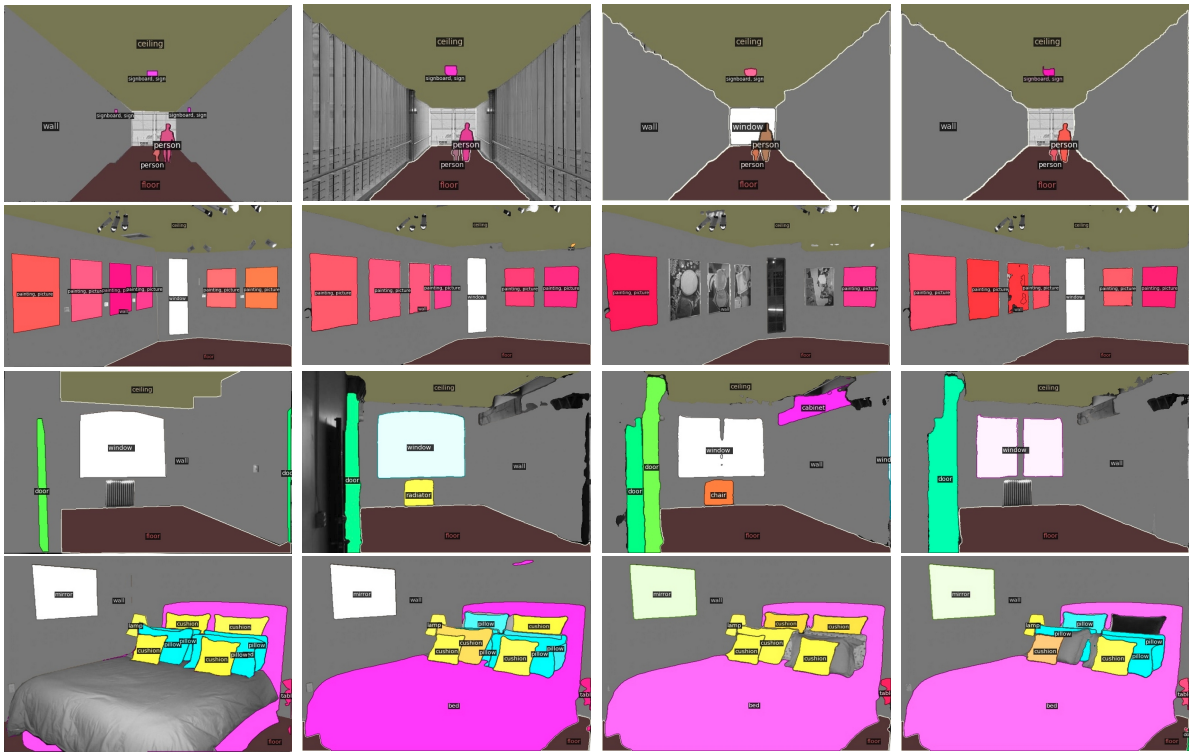
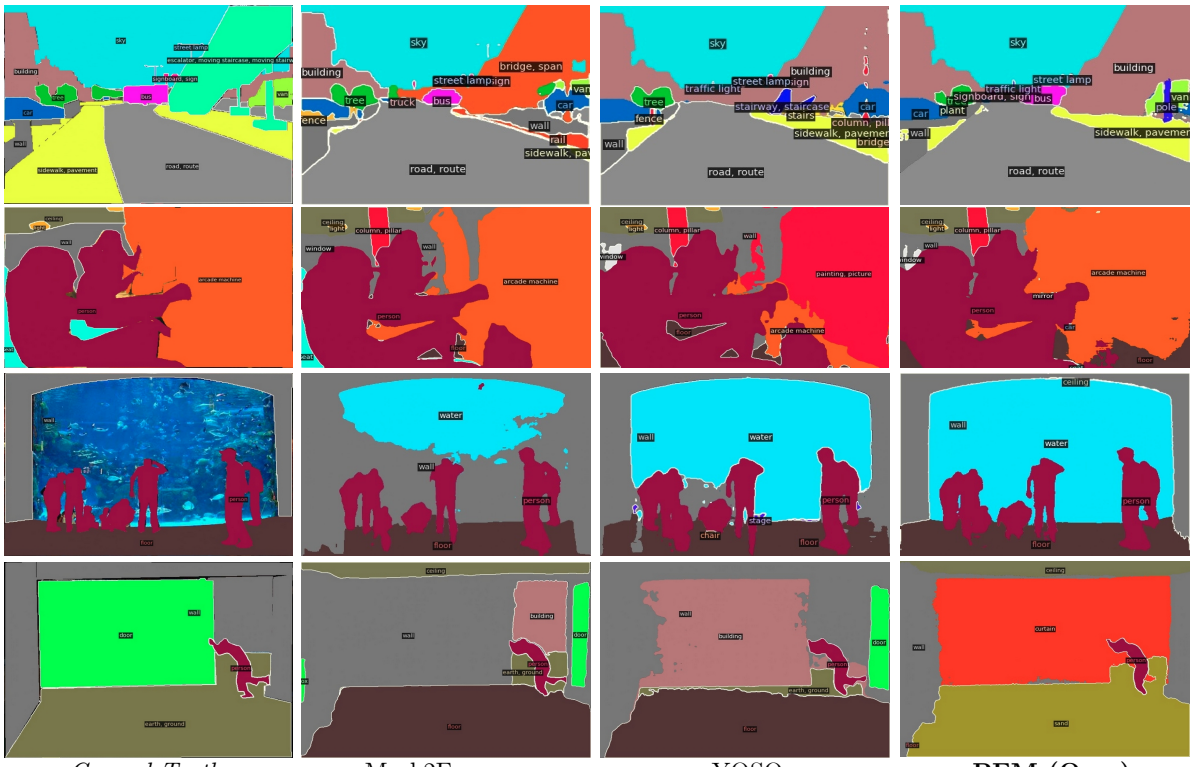


Figure 7. **Qualitative results** of PEM *v.s.* Mask2Former [5] and YOSO [18] on *semantic segmentation* on Cityscapes.



Ground Truth Mask2Former YOSO **PEM (Ours)**

Figure 8. **Qualitative results** of PEM *v.s.* Mask2Former [5] and YOSO [18] on *panoptic segmentation* on ADE20K.



Ground Truth Mask2Former YOSO **PEM (Ours)**

Figure 9. **Qualitative results** of PEM *v.s.* Mask2Former [5] and YOSO [18] on *semantic segmentation* on ADE20K.