Supplement to Large-scale Video Panoptic Segmentation in the Wild: A Benchmark

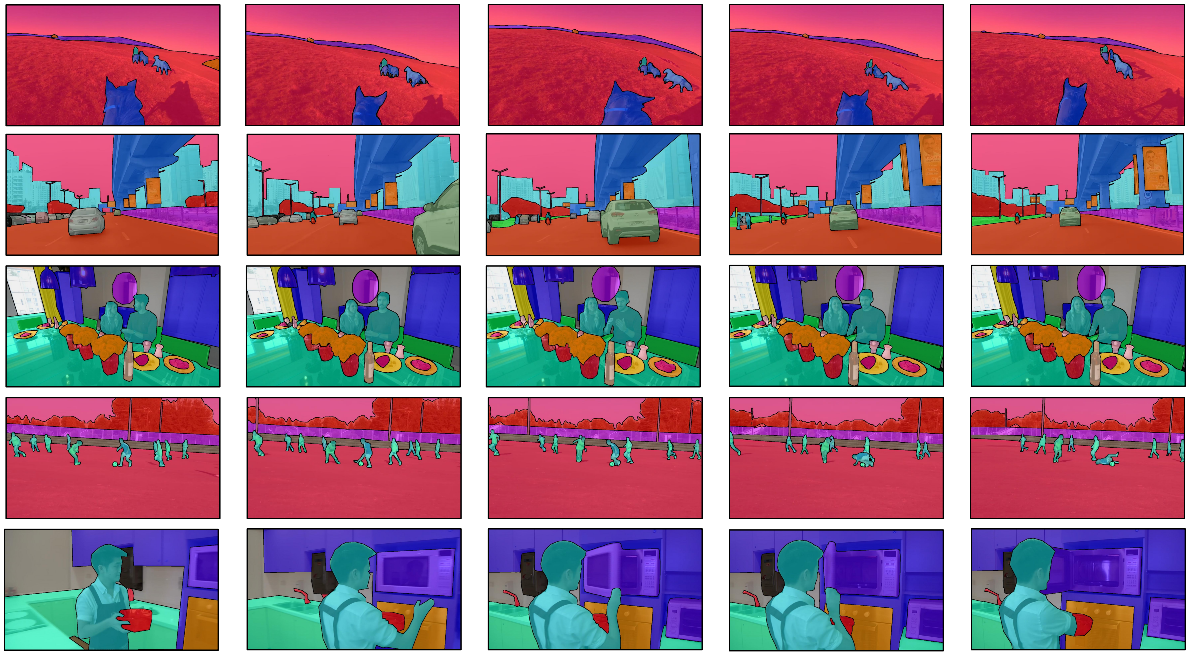

Figure 1. Examples of our large-scale VIdeo Panoptic Segmentation in the Wild (VIPSeg) dataset.

1. Implementation Details

We extend PanopticFCN [2] to our Clip-PanoFCN and the base model of PanopticFCN is pre-trained using the COCO dataset [1]. During training, the input image is augmented by random flipping, random scaling in the range of [0.8, 2.3] and random cropping to 608 × 608. We employ SGD with momentum 0.9 to optimize our model. The batch size is set to 8. We set the initial learning rate as 0.0005, weight decay as 0.0001 and the total step number as 120000. We perform the polynomial learning rate policy with factor $(1 - (\frac{iter}{iter_{max}})^{0.9})$.

2. More Dataset Statistics

Fig. 2 shows the distribution of ranked object frequencies for 124 categories. The object frequencies shows a long-tail distribution, which is typically found when a dataset is naturally collected without manual balancing. For thing-classes, "person", "chair or seat" and "car" contain the most object masks, which are common objects in the real-world. For stuff-classes, "tree", "sky" and "wall" have the most object masks.

The histograms of the number of object instances and classes per frame/video are illustrated in Fig. 4 (a) and Fig. 4 (b), respectively. We find that the two curves do not coincide with each other in the two figures, which means some instances appear or disappear during video playing, demonstrating the challenges of our VIPSeg dataset.

3. Result Visualization

Fig. 3 shows the visualization of results for our Clip-PanoFCN baseline. Since our VIPSeg is challenging with complex scenes and massive instances, the visualization results are not detailed enough especially when the instance number becomes larger. More effective methods are needed to address the challenging VIPSeg dataset.

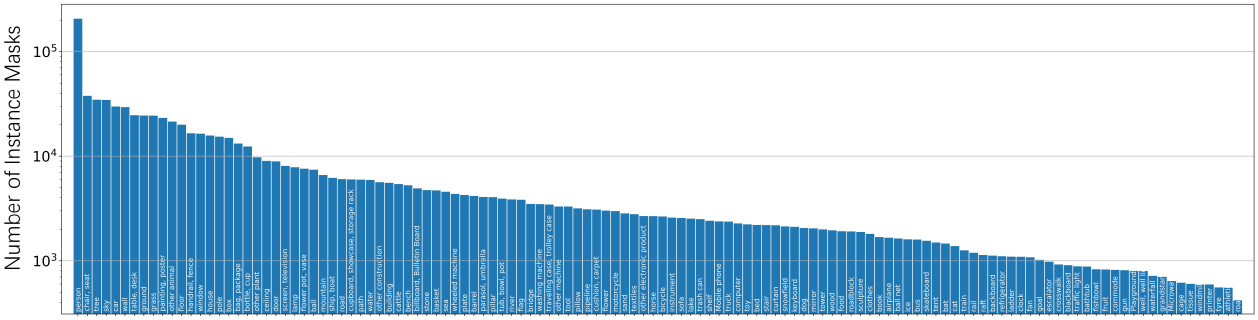

Figure 2. The histograms of the number of object instances per category.

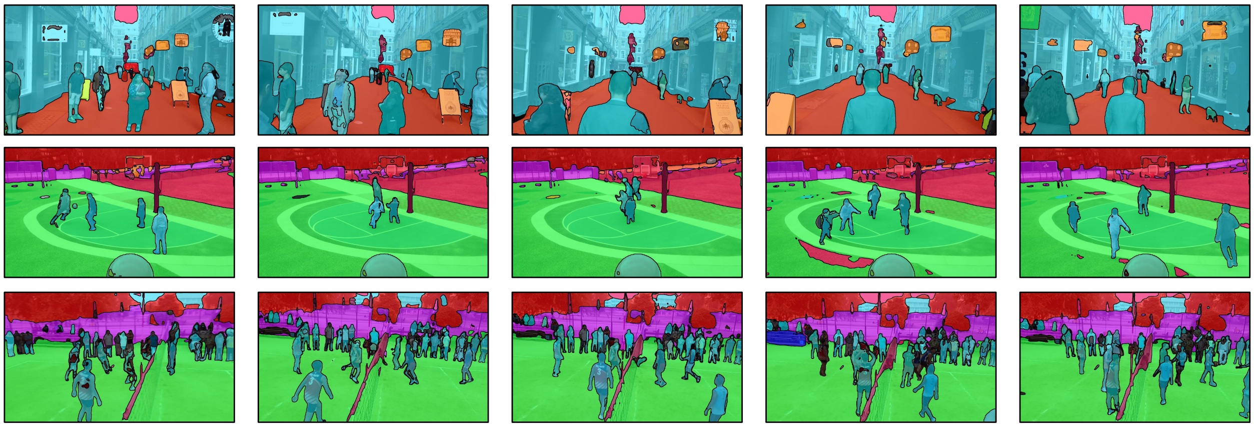

Figure 3. Visualization of results of our Clip-PanoFCN.

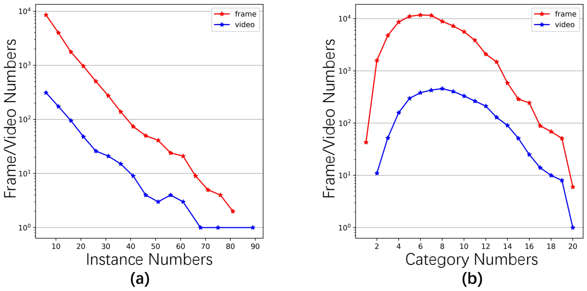

Figure 4. The histograms of the number of object instances and classes per frame/video.

4. Examples of VIPSeg

Fig. 1 shows the examples of our VIPSeg dataset. As shown in Fig. 1, our dataset contains both indoor (*e.g.*, living room, kitchen) and outdoor (*e.g.*, street view, pitch) scenes with large diversity. The variety of thing-classes in our VIPSeg is large, including person, cars, horses, ball, *etc*. All objects are carefully associated across frames.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE CVPR*, pages 1209–1218, 2018. 1

[2] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2021. 1