Global Knowledge Calibration for Fast Open-Vocabulary Segmentation

Kunyang Han[1,2] *, Yong Liu[3] *, Jun Hao Liew[4], Henghui Ding[5], Jiajun Liu[4]
Yitong Wang[4], Yansong Tang[3], Yujiu Yang[3], Jiashi Feng[4], Yao Zhao[1,2] , Yunchao Wei[1,2]

1. Text diversification

We experiment with selecting several categories and switching their names with handcrafted synonyms during inference. Note that none of these handcrafted synonyms appear in the training. For this reason, it is difficult to propose new synonyms for all categories. Therefore we try as many categories as possible. The magnitude of the degradation of mIoU after switching category names is reported below. This experiment shows that our text diversification strategy can relieve the collapse to particular known category names.

Origin	person	laptop	bicycle	airplane	truck	cat	dog	horse
↓	↓	↓	↓	↓	↓	↓	↓	↓
Synonyms	human	computer	pedal cycle	aircraft	lorry	kitty	puppy	pony
W/o TD	7.95	16.47	11.32	4.56	6.74	1.86	13.2	28.19
W/ TD	**3.10**	**15.83**	**7.88**	**1.36**	**6.16**	**-1.82**	**12.23**	**18.37**

Table 1: The magnitude of the degradation of mIoU after switching category names. The smaller the value, the better the model.

2. Video Qualitative Results

In Fig. 1, we provide some video qualitative results of our model on the VIPSeg dataset. All the videos we select contain novel categories, and our model can segment them well, showing great generalization ability in video semantic segmentation.

3. VIPSeg Dataset Split

The specific novel (unseen) categories we have defined on VIPSeg are *ice, instrument, tyre, blackboard, grandstand, fan, trash can, printer, box, basket, pillar, bathtub*. To further exhibit our seen-unseen split, in Fig. 2 we show samples of each novel category in VIPSeg. We also provide some COCO samples that contain objects belonging to the same categories, but in which these objects are annotated as ignore regions. This split can make sure our model will not learn extra knowledge during COCO training stage.

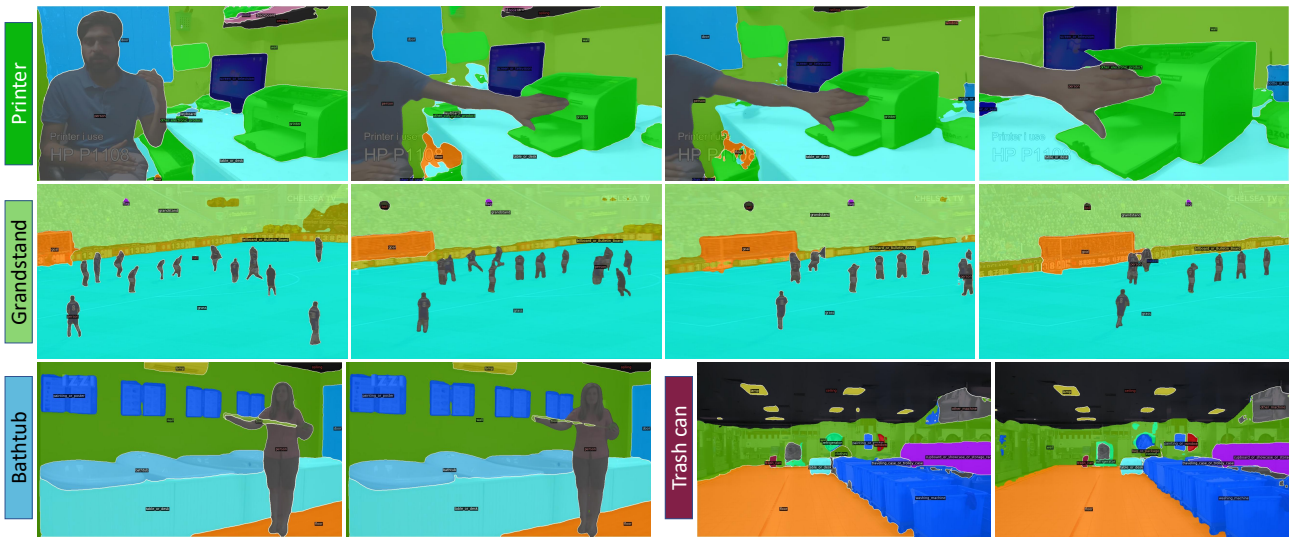

Figure 1: Qualitative results on VIPSeg dataset. We selected 4 videos, 2 or 4 frames for each video, novel category of each is shown on the left. Zoom in for more details.

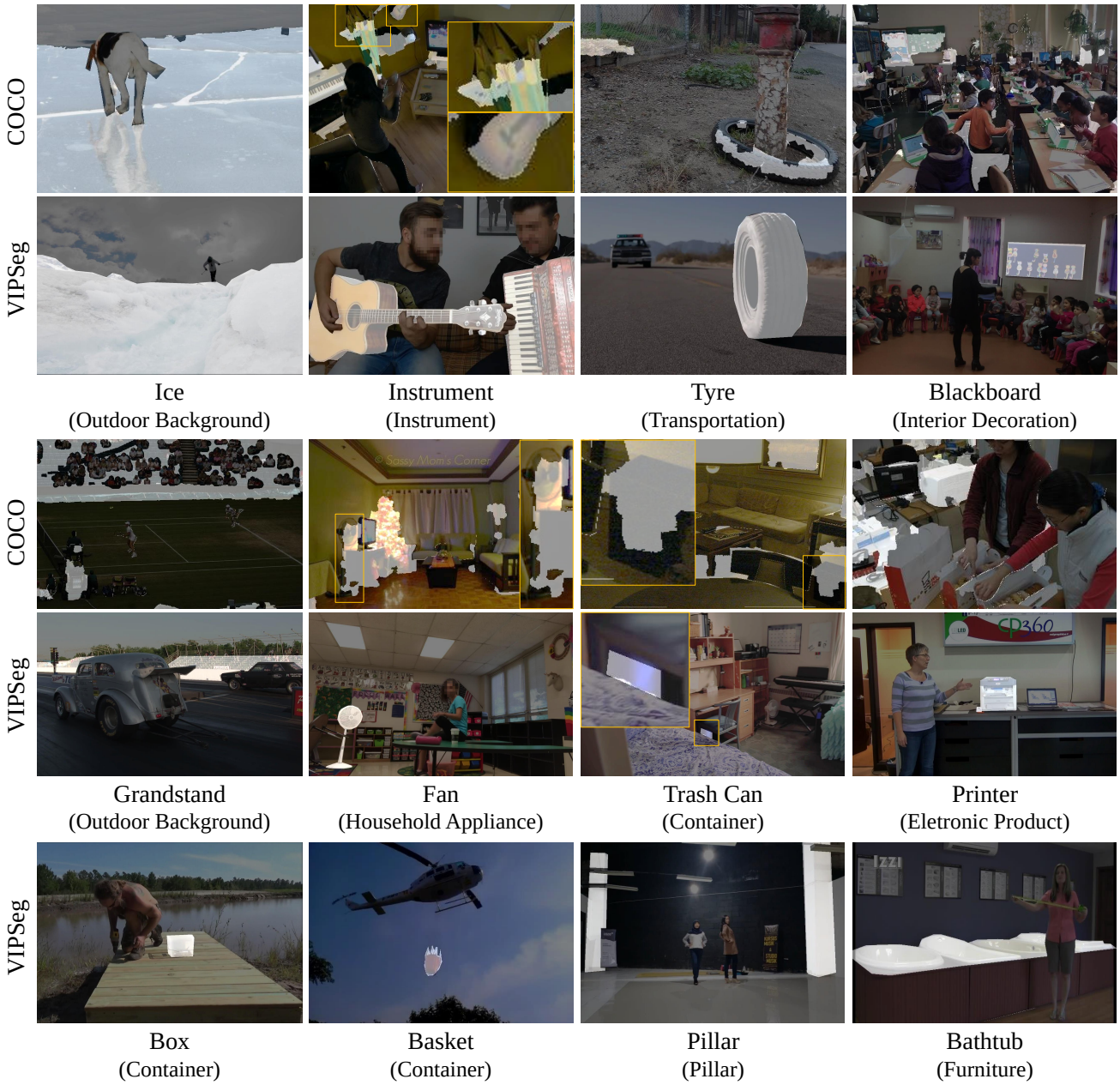

Figure 2: Samples of novel categories of VIPSeg dataset. Below the sample is the corresponding category name, and the super-category is listed in brackets. For images from VIPSeg, we highlight the regions for each category. The 1st and 3rd rows are COCO images containing objects of the same category, we highlight the corresponding regions that are annotated as "ignore" in COCO.