SegPrompt: Boosting Open-world Segmentation via Category-level Prompt Learning Supplementary Material

Muzhi Zhu[1], Hengtao Li[1], Hao Chen[1]*, Chengxiang Fan[1], Weian Mao[2,1], Chenchen Jing[1], Yifan Liu[2], Chunhua Shen[1,3]

[1] Zhejiang University, China [2] The University of Adelaide, Australia [3] Ant Group

1. Dataset Analysis

1.1. Training set

In this section, we analyze the differences in training annotations between COCO[3], LVIS[2] and our benchmark LVIS-OW, and verify the rationality and necessity of our constructed training set.

As shown in Figure 1, compared with LVIS' non-exhaustive annotations, COCO's annotations miss fewer objects which are in $\mathcal{C}_{known}$, and consequently reduce the ambiguity caused by annotations, so it is more suitable for training. However, there are obvious granularity differences between COCO and LVIS, for example, COCO only labels the complete "person", while LVIS labels the items on the person such as "skirt", "shoes", etc. Table 1 shows that if we use COCO exclusively for training, the model is completely unable to segment many common categories of LVIS, which is difficult to solve at the model and method levels. Our benchmark can alleviate this problem very well by introducing a small number of categories (64 categories, about 50% of all instances). At the same time, we remove some annotations of rare objects, making our dataset more suitable for open-world evaluation.

name	#instance	LVIS-OW	COCO
cupboard	329	53.2	3.4
polo shirt	371	54.5	9.5
sweatshirt	258	60.2	16.5
tank top(clothing)	337	43.0	4.1
billboard	270	43.2	6.3
jean	971	38.6	6.2
brake light	210	30.6	2.4
blinker	238	26.7	1.7

Table 1: **The results of models trained with different training sets on several categories of LVIS.** The LVIS-OW and COCO represent training on different training sets, and the evaluation metric is class-wise AR in Sec. 4.3 in the main text. We selected classes with a high number of occurrences on LVIS validation set (#instance $>$ 100) and a large AR gap. Many classes are completely undiscoverable relying only on COCO training.

1.2. Test set

removal ratio	#image	#class	# image per class
1%(R)	1472	337	<10
5%	5000+	544	<29
10%	10000+	688	<55
20%	20000+	846	<134

Table 2: **Detailed information on the different removal ratios.** We also show the total number of images, the total number of instances to be removed and the frequency of images for each class. The removed images are all added to the original test set lvis_v1_val, and all the removed categories are treated as unseen set $\mathcal{C}_{unseen}$.

To build a test set, a very natural question is how many categories to select for the unseen set. We designed different sizes of unseen sets according to the number of images to be removed, and the details are shown in Table 2. An intuition is that higher unseen ratios lead to more stable and convincing evaluation results, but at the same time, higher ratios lead to a reduction of the training set resulting in lower model performance. The final results are shown in the Table 3. In the case of too few unseen categories (0.1%), the evaluation results fluctuate greatly due to randomness, but when the number of unseen categories exceeds 1%, the evaluation results are gradually stable and can correctly reflect the goodness of the model, and in order not to further reduce the training data, we choose the case of 1% unseen

*HC is the corresponding author. WM was visiting Zhejiang University.

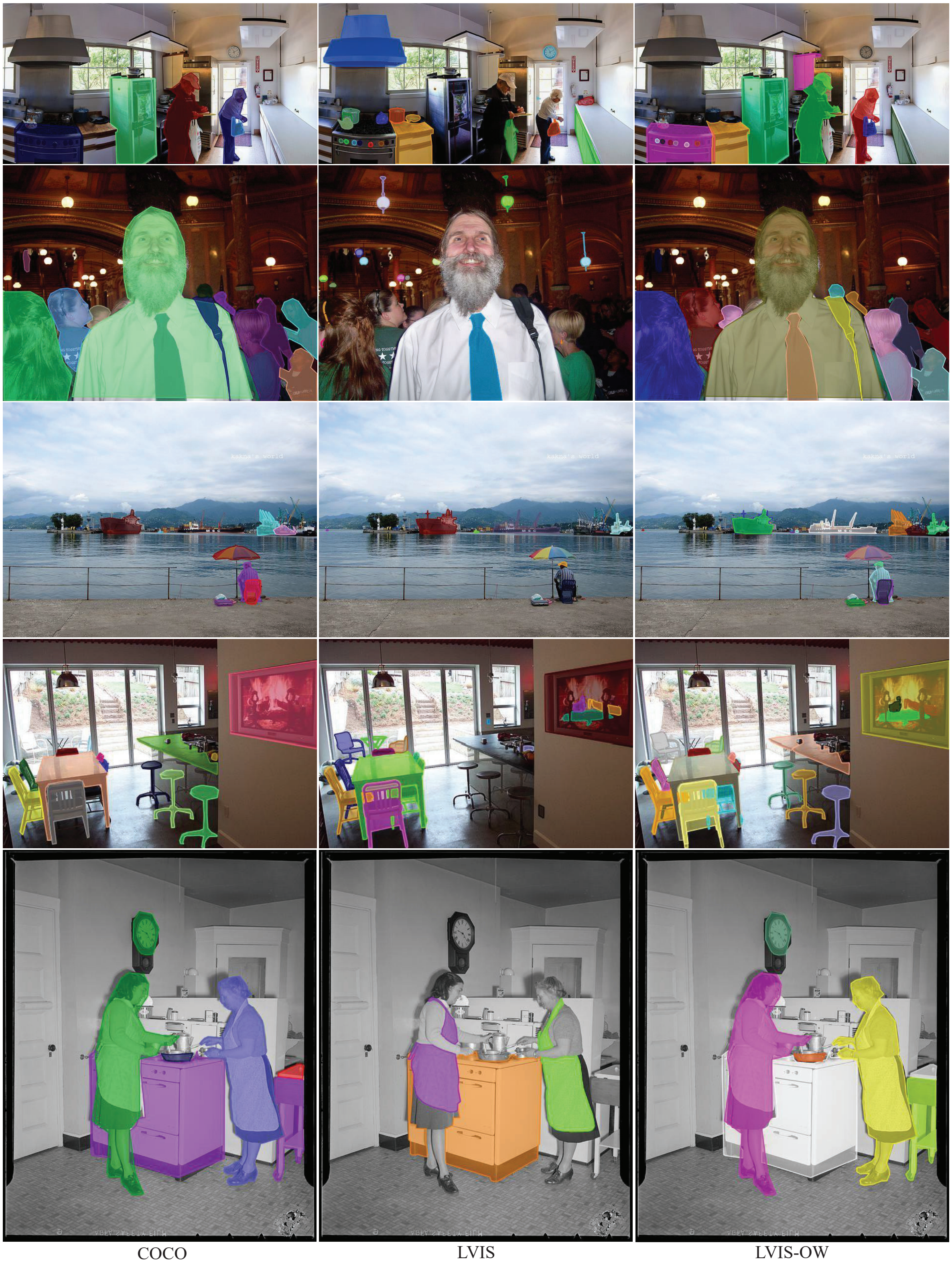

Figure 1: **Comparison of COCO, LVIS and our dataset.** COCO's annotation is more elaborate, but at a coarser granularity, and tends to annotate only instances of "person". In contrast, LVIS's annotation is very sparse but will focus on more detailed objects, for example, it tends not to segment "person", but will segment objects such as "apron", "tie", and "hat" on "person".Our benchmark combines the advantages of both and is more oriented towards an open-world setting.

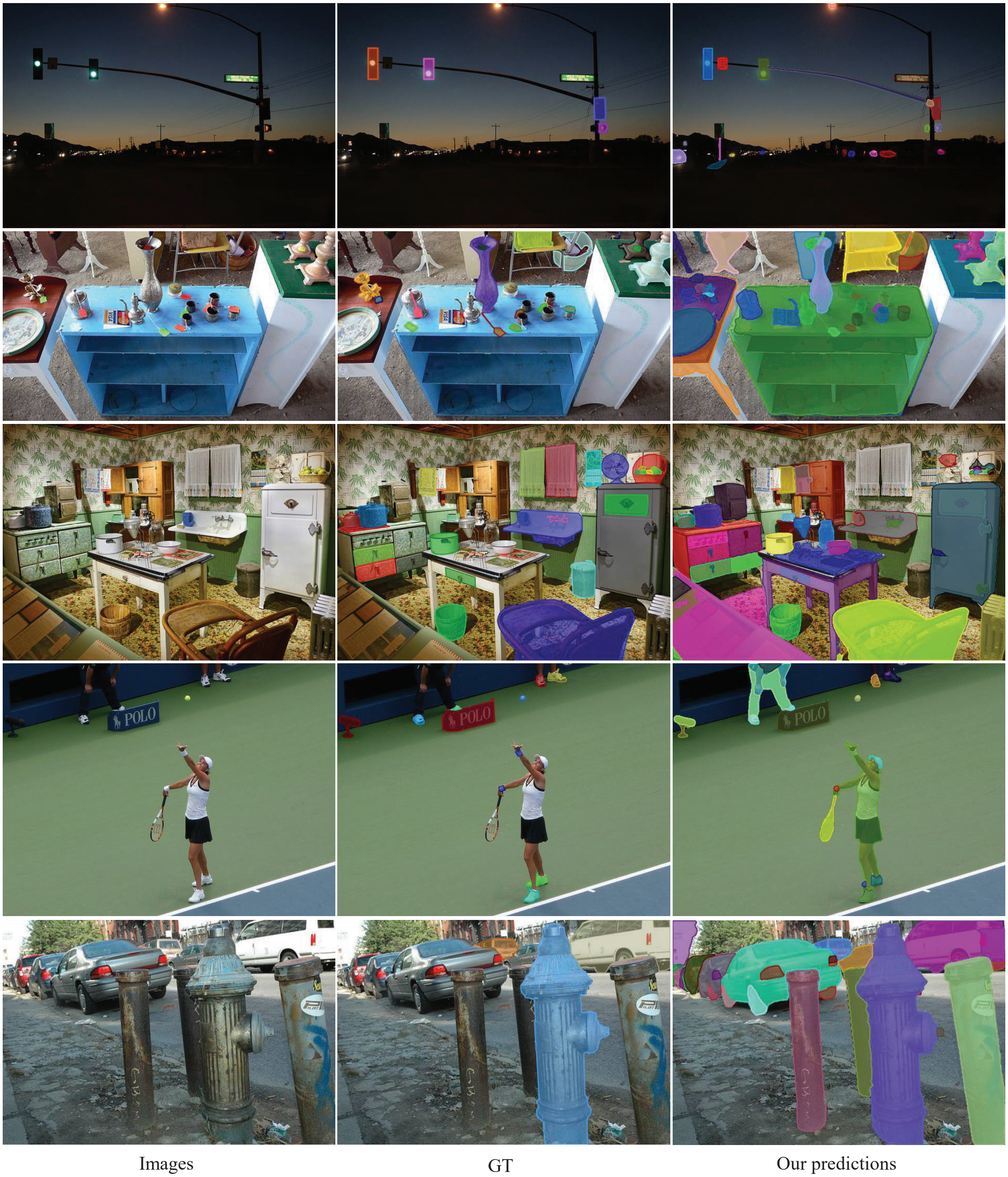

Figure 2: **Qualitative comparison of ground truth and our prediction** Our method can detect many objects that are not in the ground truth, which includes both objects belonging to unseen set $\mathcal{C}_{unseen}$ and objects that are in $\mathcal{C}_{known}$ or $\mathcal{C}_{seen}$ but are missed by the annotations.

id	AR_{all}	AR_{kn}	0.1%		1%		5%		10%		20%		origin lvis val	
			AR_{sn}	AR_{un}	AR_{sn}	AR_{un}	AR_{sn}	AR_{un}	AR_{sn}	AR_{un}	AR_{sn}	AR_{un}	AR_{sn}	AR_{un}
1	41.0	49.3	38.6	53.9	36.6	47.5	35.0	45.0	32.8	44.2	30.5	42.7	38.5	47.4
2	41.5	50.1	39.2	53.9	37.2	47.9	35.5	45.5	33.4	44.6	30.9	43.2	39.4	48.0
3	42.0	50.3	39.7	**54.3**	37.7	**48.6**	35.9	**46.2**	33.8	**45.2**	31.3	**43.8**	39.8	**47.5**
4	41.1	49.6	38.7	53.4	36.7	47.6	35.0	45.1	32.7	44.3	30.5	42.8	38.9	47.0

Table 3: **Comparison of results between different split ratios.**We train four identical models on the same training set with different random seeds, numbered 1, 2, 3, and 4. We evaluate them uniformly at the test set (lvis_v1_val + 10%) and calculate the AR according to the different divisions. For example, 0.1% means that the 100 most uncommon classes are removed as unseen set $\mathcal{C}_{unseen}$, which accounts for about 0.1% of all training sets.

categories as our experimental setup.

In Figure 2, we compare the prediction results of our method with the ground truth, and we can see that our method does detect many objects that are not in the ground truth. As we mentioned in Sec. 4.3 in the main text, if we use AP as a metric, these additional detected objects will be counted as false positives, resulting in extremely low AP, so AR is a more appropriate metric.

2. Additional Ablations

In the additional ablation experiments, we don't use example supervision, and the other settings are the same as in the main text.

Training iterations. Since our model is initialized with the already trained class-agnostic Mask2former before additional training, the experiment in Table 4 is conducted in order to exclude the possibility that the performance gain comes only from the additional training iterations. Additional training based on the original mask2former does not effectively improve the model's open-world oriented segmentation ability, instead our method further improves performance with more training iterations.

The maximum number of categories C_{max}. The maximum number of categories C_{max} that can be predicted per image is a critical hyperparameter, because the number of negative class C_{neg} is tightly relative to C_{max}, for instance, a large C_{max} can bring a lot of negative class. Thus, we perform an experiment to investigate C_{max}, as shown in Table 5. Notably, the number of the maximum prediction per category K is reduced when C_{max} increases, because the number of queries N^{query} is fixed to 300 and $N^{query} = C_{max} \times K$. The experiment results show that the model performance is improved steadily when C_{max} is reduced. This trend reaches a state of saturation when $C_{max} = 15$.

	Iters	AR_{all}	AR_{kn}	AR_{sn}	AR_{un}	AR_s	AR_m	AR_l
M2F	0	42.5	49.9	39.0	47.4	27.2	51.5	68.1
M2F	90k	42.1	49.8	38.5	47.1	27.0	50.9	67.5
Ours	90k	**44.1**	**51.9**	**40.6**	**49.1**	**28.9**	53.3	**70.6**

Table 4: **The results of different training iterations.** As the number of iterations increases the performance of our model can be further improved. The result in the first row is the Mask2former for 50 epochs of class-agnostic training, which is also the initialization weight of the class-agnostic baseline branch of our model. The second row shows that more training based only on the class-agnostic branch of Mask2former[1] does not improve the model and even degrades the performance. The last line reflects the effectiveness of our method.

C_{max}	K	AR_{all}	AR_{kn}	AR_{se}	AR_{un}	AR_s	AR_m	AR_l
300	1	43.0	51.2	39.4	48.1	27.3	52.6	69.6
150	2	43.8	50.9	40.5	48.5	28.0	53.3	70.9
75	4	44.0	51.2	40.6	48.8	27.9	53.8	**71.2**
50	6	44.2	51.6	**40.8**	49.1	28.5	53.9	70.5
30	10	44.1	51.9	40.6	49.1	28.9	53.3	70.6
20	15	**44.4**	52.1	**40.8**	49.5	**29.0**	**54.0**	70.6
15	20	**44.4**	**52.2**	40.6	**50.0**	**29.0**	53.9	70.3

Table 5: **Varying the maximum number of categories C_{max}.** C_{max} is the maximum number of categories that can be predicted per image and K is the number of the maximum prediction per category. The number of queries N^{query} in the prompt-based prediction branch equals C_{max} multiply K. For a fair comparison, N^{query} is fixed to 300 while the C_{max} is varied.

3. Implementation Details

Our method can be trained from scratch or based on a pre-trained Mask2former. Except for the experiments on the full LVIS training set and COCO→ LVIS, other experiments are based on the latter, and we find that the effect of freezing the backbone is better.

Prompt learning mechanism. For prompt extraction and learning, we filter out the mask annotations with an area smaller than 100 on 1024×1024 images. This is caused by the mechanism of mask-attention. The area of the small mask on the smaller feature map will become 0, forcing the mask-attention to focus on all regions of the whole picture, which cannot effectively extract the corresponding object information, resulting in the quality of the prompt cache being poor.

Open-vocabulary/Few-shot Segmentation. Unlike the above, where the prompt extraction branch and the prompt prediction branch are used as auxiliary training modules, here we need to use the prompt prediction branch directly for inference, and the class-specific prompt will use the reference attention to combine the information from the class-agnostic branch to segment the objects in the image corresponding to the class. In addition, there are some differences in the training process, we introduce the image-level label of the seen $\mathcal{C}_{seen}$ into the prompt prediction branch, and the candidate query of this part only receives the supervision of the classification loss, but not the supervision of the mask. In order to make each category query more accurate in locating objects in the corresponding category and to facilitate the constraint by classification, we set K to 1, i.e., each category has only one query for prediction. For open-vocabulary segmentation, we replace all learnable class-specific embedding $\mathbf{s}$ with fixed CLIP[4] text embedding and drop the original prompt extraction branch, and the model can do open-vocabulary segmentation after training. For the few-shot segmentation, we keep the prompt extraction branch. A small number of support images are used as input, and the model also performs a momentum update on the prompt cache and then uses the prompt to segment the test images.

References

[1] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 1290–1299, 2022.

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5356–5364, 2019.

[3] 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 *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021.