Open-Vocabulary Semantic Segmentation with Decoupled One-Pass Network –Supplementary Material–

Cong Han[1*] Yujie Zhong[1*] Dengjie Li[1] Kai Han[2†] Lin Ma[1†]
[1]Meituan Inc. [2]The University of Hong Kong
hancong0911@163.com jaszhong@hotmail.com kaihanx@hku.hk

1. Class-agnostic Mask Proposal Network

We provide a detailed illustration of the mask proposal network in this section. We utilize a modified MaskFormer as our mask proposal network, which is mentioned in 3.1 of the main paper. We present the overall framework of our class-agnostic mask proposal network in Figure S1. The model architecture is essentially the same as MaskFormer, while the classification branch is removed. We use a backbone to extract image features F_B. The image features are fed into a pixel decoder and a transformer decoder to generate N mask embeddings and per-pixel embeddings F_P. Finally, we combine the N mask embeddings and F_P using matrix multiplication to get masks M. The quality of the proposal masks is assessed in Section 3.2.

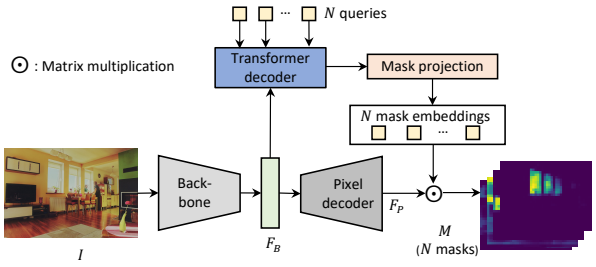

Figure S1: The architecture of class-agnostic mask proposal network.

2. Visual Prompt Learning

As discussed in Section 3.2 in the main paper, we improve the performance of the baseline method by fine-tuning the pre-trained model with prompt learning. The text prompt learning is explained in detail in the main paper, so here we elaborate on the visual prompt learning that we adopted. The architecture of visual prompt learning is demonstrated in Figure S2. We compare two forms of visual prompt learning: prepending prompts and adding prompts. The difference between these two forms lies in how the prompts are combined with image patch embeddings. *Prepending prompts* refer to prepending P prompt embeddings before O image patch embeddings, resulting in $P + O$ embeddings, together with a class embedding. *Adding prompts* involves adding a prompt embedding to each image patch embedding in an element-wise manner, and the length of prompt embeddings should be the same as the length of image patch embeddings.

In this section, we first provide the details of the datasets and evaluation metrics, and then provide further analysis of our methods, by including the ablation study on different prompt learning approaches, the effectiveness analysis (of the main components) on Pascal VOC dataset, and the performance of the mask proposal network.

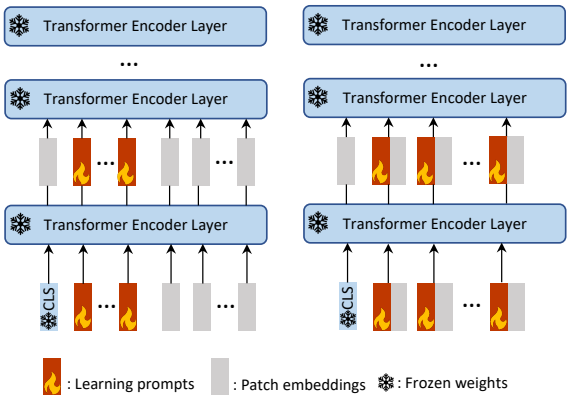

Figure S2: Overview of visual prompt tuning. There are two implementations: prepending prompts (**Left**) and adding prompts (**Right**).

2.1. Datasets and Evaluation Metrics

Datasets. **COCO-Stuff** is a large dataset for semantic segmentation that span over 171 categories including 80 things, 91 stuff. It contains 117k training images and 5k validation images. **PASCAL VOC** contains 11,185 training images and 1,449 validation images from 20 classes. **PASCAL Context** is a set of additional annotations for PAS-

*Equal contribution.
†Corresponding authors.

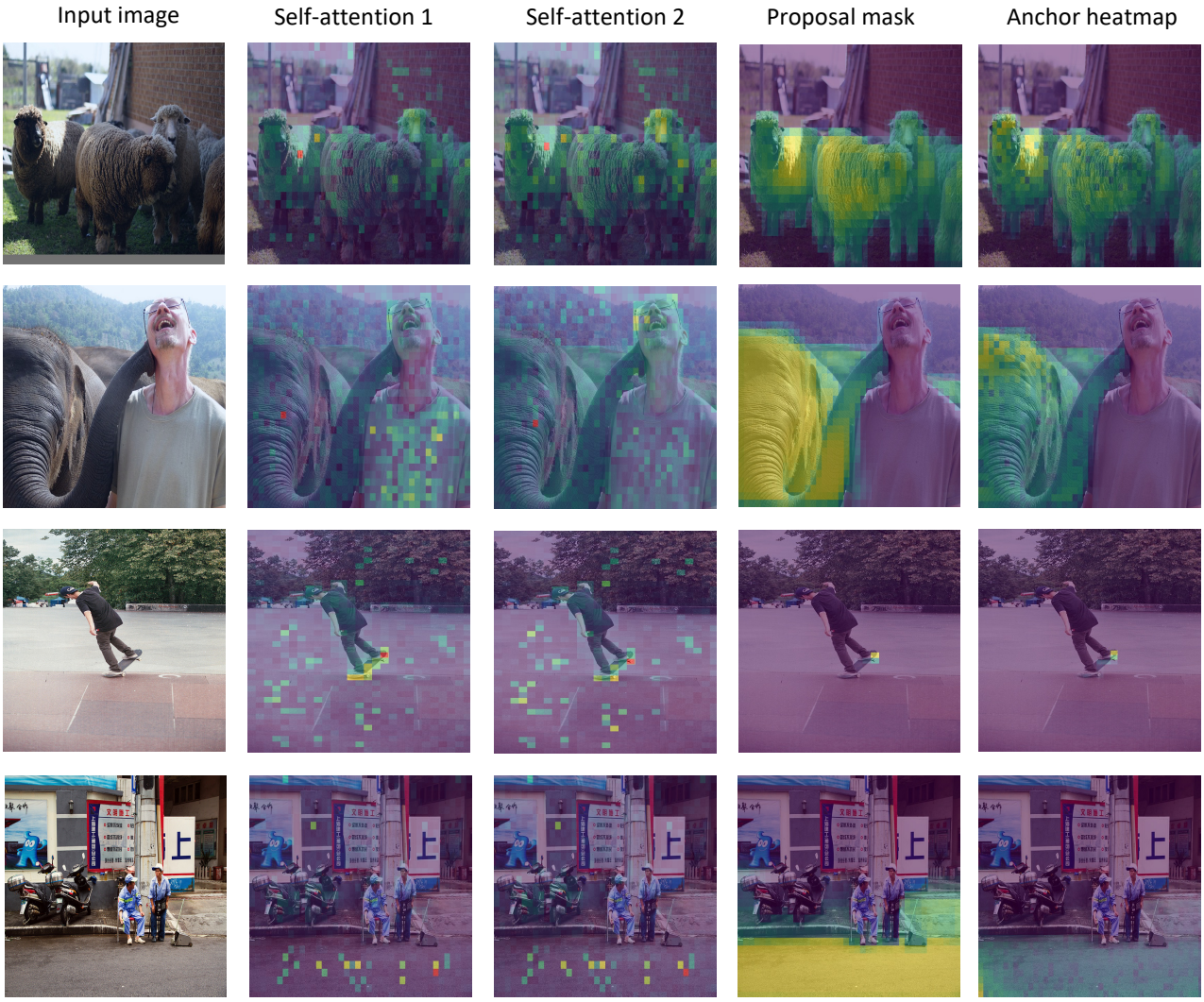

Figure S3: The visualization of the heatmaps. The three categories in the four rows are: sheep (seen), elephant (seen), skateboard (**unseen**) and road (**unseen**).

CAL VOC 2010. It contains 4,998 training images and 5,005 validation images. We select a subset of 59 frequent classes for use. **ADE20K** contains more than 20K scene-centric images for training and 2k images for validation. There are totally 150 semantic categories, which include stuff and discrete objects.

3. Experiment

Data split. We choose two types of data splits for validating our method on zero-shot semantic segmentation (ZS3) setting and cross-dataset setting respectively. For ZS3 setting, we follow the class split in [2]. In particular, on COCO-Stuff, we choose 156 classes as the seen classes and the rest 15 classes as the unseen testing classes. On Pascal VOC 2012, we choose 15 classes as the seen classes and the rest 5 classes as the unseen testing classes. For cross-dataset setting, we train the model on COCO-Stuff seen classes dataset and validate on other datasets.

Evaluation metrics. Following the previous work, we measure pixel-wise classification accuracy (pAcc) and mean IoU (mIoU) for seen and unseen classes denoted as mIoU(S) and mIoU(U) respectively. Additionally, we compute the harmonic mean IoU (hIoU) among seen and unseen classes by the previous works [36], which is calculated as $hIoU = \frac{2*mIoU(S)*mIoU(U)}{mIoU(S)+mIoU(U}$. For cross-dataset validation, we use mIoU as the evaluation metric.

3.1. Prompt Learning

Analysis on prompt learning. In Table S1, we evaluate the effectiveness of different prompt learning approaches

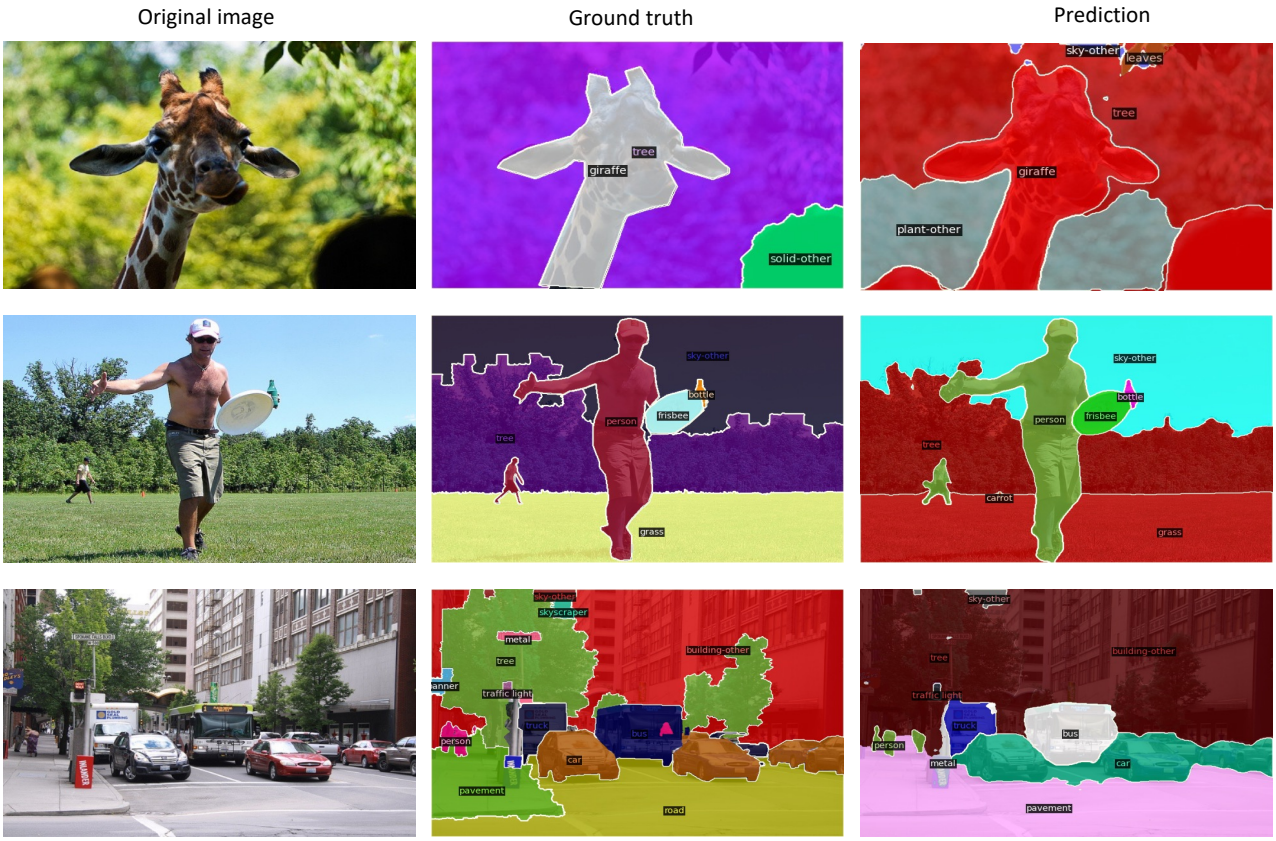

Figure S4: The visualization of semantic segmentation. From left to right: origin images, ground truth semantic segmentation maps, and predictions.

Type	End-to-end	pAcc	hIoU	Seen	**Unseen**
Text	✗	15.6	7.0	6.2	8.0
Text	✓	25.7	7.2	7.6	6.9
Vision A.	✓	22.6	6.4	6.3	6.5
Vision P.	✓	16.7	6.7	6.4	7.4
Both	✓	**27.6**	**8.3**	**8.0**	**8.7**

Table S1: Ablation study on Prompt Learning. $Vision\ P.$ and $Vision\ A.$ mean prepending and adding prompts for visual prompt tuning. $Both$ refers to combining end-to-end text prompt learning and adding visual prompt learning.

Method	All	Seen	**Unseen**
Recall@30	0.59	0.57	0.71
Recall@50	0.45	0.44	0.56

Table S2: Recall of mask proposals on COCO-Stuff dataset. $Recall@30$ and $Recall@50$ means recall at IoU 30% and 50% respectively.

for segmentation task building upon the baseline method. In Section 3.2 of the main paper, we propose improving the baseline method with prompt learning. To evaluate the impact of prompt learning, we conduct experiments with different types of prompts, based on the baseline method. We can observe that all prompt learning approaches can enhance model performance. The best results are achieved when combining end-to-end text prompt learning with adding visual prompt.

Method	layers	hIoU	Seen	**Unseen**
Baseline	-	7.0	6.2	8.0
GPS	[1]	4.6	4.5	4.7
GPS	[1-6]	-	0.0	0.1
GPS	[10-12]	1.9	1.3	3.3

Table S3: Ablation study on generalized patch severance.

3.2. Generalization of Class-agnostic Mask Proposal Network

To evaluate the generalization of the mask proposal network which is only trained on images belonging to seen classes, we report the recall of mask on COCO-Stuff dataset in Table S2. We calculated two metrics, $Recall@30$ and $Recall@50$. The results indicate that the mask proposal network can provide class-agnostic masks for both seen and unseen classes. The network exhibits a satisfactory level of

generalization performance.

4. Visualization

Heatmap visualization. We show the visualization of heatmaps in Figure S3 which are supplementary cases for Section 5.3 of the main paper. The self-attention in the transformer encoder layer retrieves information from a global scope, which may actually introduce more noise, and is therefore disadvantageous for the segmentation task. $sheep$ class, $elephant$ class and $skateboard$ class are included in "thing" category, and $road$ are categorized as "stuff" category. It demonstrates that our method can work well on both "thing" and "stuff" categories. CAL module is capable of guiding the model to focus on more distinctive regions, which can enhance its ability to accurately classify different categories.

Segmentation results. As shown in Figure S4, we present the visualized predictions of DeOP. Our method can effectively segment regions belonging to different categories. The method can achieve impressive results even when presented with regions of unseen categories (*giraffe*, *tree*, *frisbee* and *grass*), indicating its remarkable generalization performance and effectiveness for the zero-shot segmentation task.