Supplementary for Understanding Personal Concept in Open-Vocabulary Semantic Segmentation

Sunghyun Park* Jungsoo Lee* Shubhankar Borse Munawar Hayat
Sungha Choi† Kyuwoong Hwang Fatih Porikli
Qualcomm AI Research‡
{sunpar, jungsool, sborse, hayat, sunghac, kyuwoong, fporikli}@qti.qualcomm.com

Figure 1. Examples of segmentation maps we used in ADE^{per} dataset. We combine the open-vocabulary segmentation map and the foreground segmentation map. Specifically, we annotate the class we want to personalize as the '<my concept>' category index instead of the original category index (*e.g.* chest and hot tub classes)

A. Implementation Details

A.1. Further details

In this section, we describe further implementation details of experiments. For both SAN [4] and ODISE [3], we use the batch size of 1, 100 number of masks, and $\lambda_{\text{Z}}^{\text{neg}} = 0.1$ across all datasets. Also, regarding the injection of visual embedding, we set the $\alpha = 0.1$ for FSS^{per} and CUB^{per} and $\alpha = 0.01$ for ADE^{per} for both models. For SAN, we use the learning rate of 5e-4 and set $\lambda_{\text{M}}^{\text{neg}} = 10$ for FSS^{per} and ADE^{per} while using $\lambda_{\text{M}}^{\text{neg}} = 500$ for CUB^{per}. For ODISE, we use the learning rate of 2e-3 for FSS^{per} and 1e-4 for CUB^{per} and ADE^{per}. We set $\lambda_{\text{M}}^{\text{neg}}$ =10.0, 500.0, and 1.0 for FSS^{per}, CUB^{per}, and ADE^{per}, respectively. We use CLIP [1] and Stable Diffusion v1.3 [2] for injecting visual embedding for SAN and ODISE, respectively. The total number of trainable parameters of our method is 0.4M, which denotes that our method requires a negligible amount of additional parameters for personalization. We conduct all experiments using one A5000 GPU with less than 24GB of GPU memory usage.

A.2. Dataset Preprocessing

Fig. 1 shows the segmentation maps we utilized. In order to personalize the open-vocabulary segmentation models, our method only requires the foreground segmentation maps for training. In other words, during the personalization or training process, it is necessary to provide annotations only for the segments that the user is interested in. For the quantitative evaluation in our experiments, we employ combined

*These two authors contributed equally to this work. †Project lead. ‡Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

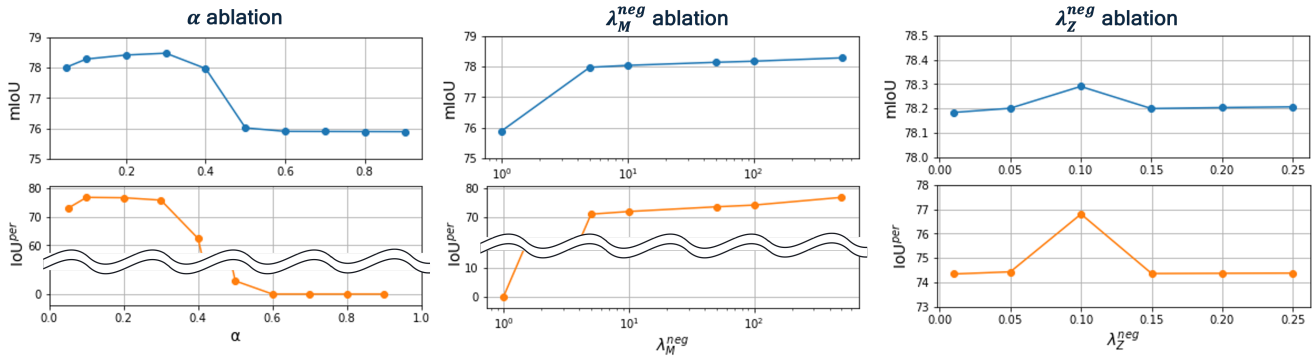

Figure 2. Hyperparameter ablation studies on CUB-200 dataset.

segmentation maps, as shown in Fig. 1. Unlike the previously used open-vocabulary segmentation maps, we annotate categories specified as visual concepts with '¡special¿' category index for quantitative evaluation. However, different from ADE^{per}, the FSS^{per} and CUB^{per} datasets do not include open-vocabulary segmentation maps and only contain foreground segmentation maps. Consequently, for the quantitative evaluation, we use predictions of SAN and ODISE for the open-vocabulary segmentation maps. Furthermore, the vocabulary set used for the open-vocabulary segmentation models was based on the categories from the COCO-stuff dataset. Previously, there was no established evaluation protocol capable of accurately measuring the performance of open-vocabulary segmentation and personal visual concepts quantitatively. Therefore, we develop combined segmentation maps that enable the concurrent measurement of IoU^{per} and mIoU.

The FSS-1000 and CUB-200 datasets exhibit high visual similarity within a single category and include foreground masks, making them suitable for assessing personalization performance for specific visual concepts. The ADE-20K-847 dataset has been extensively used for the quantitative evaluation of existing open-vocabulary segmentation models. Thus, we utilize a subset of ADE-20K to construct ADE^{per} for our experiments. As described in the main paper, we intentionally include an equal number of images with and without personal visual concepts since models that are fine-tuned to recognize personal concepts overconfidently predict other concepts as personal concepts.

For the CUB^{per} dataset, all 200 classes are used for evaluation, whereas for the FSS^{per} and ADE^{per} datasets, only 30 classes are selected for the experiments, as described in the main paper. From the FSS-1000 and ADE-20K-847 datasets, we select categories that are challenging to recognize based solely on text descriptions, are rare, or can be easily confused with other categories. In the FSS dataset, we create pairs of similar categories and use images from these similar yet distinct categories as images without the personal concept (*e.g.* wooden boat−banana boat). Table 1 describes the categories we selected for FSS^{per} and ADE^{per} datasets.

Dataset	Categories
FSS^{per}	adidas logo1−jordan logo / apple icon−yonex icon / banana boat−wooden boat / bath ball−pokermon ball / bulbul bird−chickadee bird / cactus ball−gym ball / croquet ball−french ball / esport chair−ganeva chair / folding chair−hair razor / golf ball−soccer ball / hair drier−rocking chair / jay bird−magpie bird / kappa logo−nike logo / kobe logo−puma logo / ping-pong ball−rugby ball
ADE^{per}	altarpiece / ashtray / banner / beacon / booklet / candelabrum / canister / chest / console table / crane / dirt track / easel / embankment / footbridge / hot tub / hovel / kettle / kitchen island / pane / parking meter / pier / place mat / postbox / rod / runway / saucepan / shower stall / soap dispenser / stretcher / towel rack

Table 1. The categories we selected for FSS^{per} and ADE^{per}.

B. Additional Experiments

B.1. Hyper-parameter Sensitivity

Fig. 2 demonstrates the sensitivity of the hyper-parameters used in our work. Our proposed method includes the following hyper-parameters: 1) interpolation value between visual and textual embeddings denoted as α, lambda values for $\mathcal{L}_{\text{M}}^{\text{neg}}$ and $\mathcal{L}_{\text{Z}}^{\text{neg}}$ denoted as $\lambda_{\text{M}}^{\text{neg}}$, and $\lambda_{\text{Z}}^{\text{neg}}$, respectively. For the model and dataset in the experiments, we used SAN and CUB-200, respectively. Regarding α, we empirically found that we achieve promising performances when α is set to

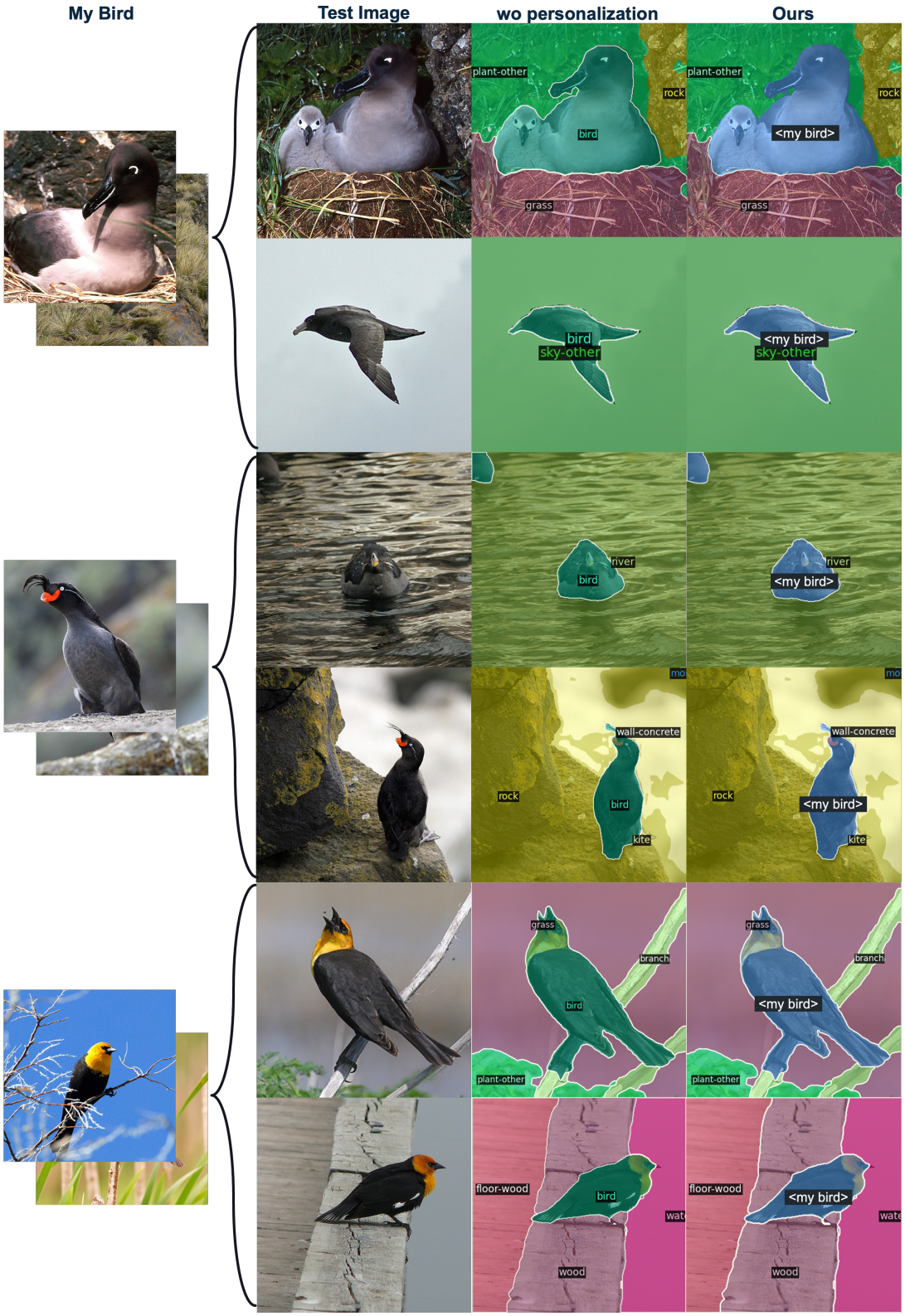

Figure 3. Additional segmentation results on CUB^{per}. While SAN wihtout personalization fails to capture the personal visual concept (*i.e.* my bird), our method applied to SAN recognizes it.

Dataset	Method	IoU^{per}				mIoU			
		$K=1$	$K=3$	$K=5$	Avg.	$K=1$	$K=3$	$K=5$	Avg.
FSS^{per}	ODISE [3]	10.69	10.69	10.69	10.69	**23.86**	**23.86**	**23.86**	**23.86**
	+ Visual Embedding	0.00	0.00	0.00	0.00	23.37	23.37	23.37	23.37
	+ Ours	**30.97**	**33.11**	**34.05**	**32.71**	21.83	22.68	22.94	22.48
	SAN [4]	41.08	41.08	41.08	41.08	55.68	55.68	55.68	55.68
	+ Visual Embedding	0.00	0.00	0.00	0.00	56.12	55.36	54.45	55.31
	+ Ours	**49.80**	**54.09**	**56.80**	**53.56**	**56.40**	**56.73**	**55.85**	**56.32**
CUB^{per}	ODISE [3]	0.02	0.02	0.02	0.02	**47.48**	**47.48**	**47.48**	**47.48**
	+ Visual Embedding	0.02	0.02	0.02	0.02	47.71	47.36	47.31	47.46
	+ Ours	**5.39**	**5.90**	**5.99**	**5.76**	45.16	44.94	44.88	44.99
	SAN [4]	68.25	68.25	68.25	68.25	77.32	77.32	77.32	77.32
	+ Visual Embedding	0.70	0.00	0.00	0.23	72.77	67.13	65.98	68.63
	+ Ours	**76.70**	**77.21**	**76.80**	**76.90**	**77.36**	**77.85**	**78.29**	**77.83**

Table 2. Comparisons with baseline using visual embedding. We apply our method on both SAN [4] and ODISE [3] on FSS^{per} and CUB^{per}. We vary K, the number of images and masks, to 1, 3, and 5.

low values. This indicates that while injecting visual information indeed improves performances, the information of textual embedding need to be included more than that of visual embedding. We select $\alpha = 0.1$ since it achives the best IoU^{per}. Also, the performances saturate as $\lambda_{\text{M}}^{\text{neg}}$ reaches to a certain value, so we select ($\lambda_{\text{M}}^{\text{neg}} = 500$) when further performance gain is no longer observed. Additionally, using different values of $\lambda_{\text{Z}}^{\text{neg}}$ shows consistent performances, so we select the value ($\lambda_{\text{Z}}^{\text{neg}} = 0.1$) with the best performance.

B.2. Segmentation Results

Fig. 3 compares the segmentation results of SAN without personalization and our method applied to SAN using CUB^{per} as the dataset, which demonstrate the effectiveness of understanding personal visual concept (*i.e.* my bird). For SAN without personalization, we provide the text descriptions of the visual concept we want to personalize. Such a qualitative analysis clearly demonstrates that our newly proposed task, personalized open-vocabulary semantic segmentation, is challenging with the existing open-vocabulary segmentation model, and it needs to be explored. We believe that our study serves as a cornerstone to further improve performance on personalized open-vocabulary semantic segmentation task.

B.3. Comparison with Visual Embedding

Since our work is the first to propose the personalized open-vocabulary semantic segmentation task, we lacked baseline models for comparison. In order to further demonstrate the effectiveness of our proposed approach, we compared our method against a baseline that uses masked visual embeddings instead of text embeddings. Specifically, given few training images and masks corresponding to a given personal concept, we extracted visual embeddings using the image encoder of CLIP and replaced the text embeddings of personal concept with the averaged visual embeddings. During the experiment, we maintained the other components of open-vocabulary segmentation models (*e.g.*, SAN, ODISE).

Dataset	Method	IoU^{per}	mIoU
FSS^{per}	SAN [4]	28.59	57.66
	+ Ours	**38.47**	**60.09**
CUB^{per}	SAN [4]	24.25	81.62
	+ Ours	**40.63**	**81.71**

Table 3. Quantative results on **concat** datasets. We apply our method on SAN [4] on FSS^{per} and CUB^{per}. K is set to 5.

Table 2 shows that our method significantly outperforms the baseline using masked visual embeddings on the FSS^{per} and CUB^{per} datasets. This result demonstrates the necessity of using the representation space of text encoders for tasks related to open-vocabulary segmentation. Similar results are observed in the α ablation study in Fig. 2, where the performance in IoU^{per} drops dramatically as the proportion of masked visual embeddings increases excessively.

B.4. Further Experiments on Distinguishing Personal Concept from Similar Classes

As shown in the qualitative results of the main paper (Fig. 3, 5, and 6), our method can distinguish between the target visual concept (*e.g.*, "my boat", "my bird") and its corresponding similar classes (*e.g.*, "boat", "bird") within the same image. For the results, we used FSS and CUB datasets

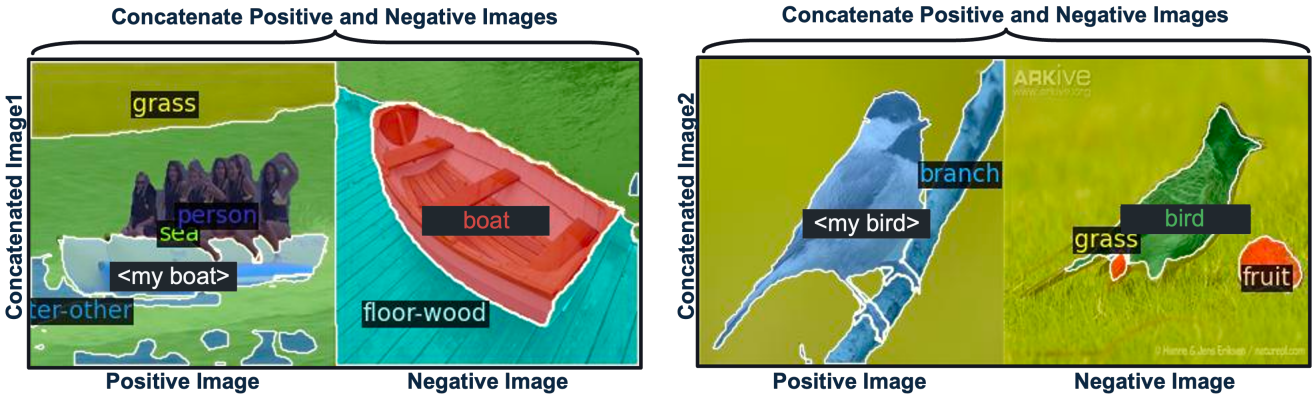

Figure 4. Test images of **concat** dataset (FSS^{per}). These datasets are used for evaluating the performance on distinguishing between the target visual concept (*e.g.*, "my boat", "my bird") and its corresponding similar classes (*e.g.*, "boat", "bird") within the same image.

annotated with fine-grained segmentation maps, which include several images and masks on specific fine-grained classes. However, we found that these datasets often contain only one object per image, which limits our evaluation on distinguishing the personal concept with its corresponding similar classes within the same image.

To this end, we conducted an additional experiment by concatenating two images horizontally: 1) a positive image that contains the target visual concept and 2) a negative image that contains the similar class but without the target visual concept. Then, we treated the concatenated images as a single image and input it for segmentation. For better understanding, we show examples of concatenated positive and negative class images, which we refer to as the "concat dataset" in Fig. 4. Table 3 demonstrates that our approach significantly improves IoU^{per} compared to existing open-vocabulary segmentation models in such an experimental setting. These results show that our method can effectively distinguish between the target visual concept and similar but different classes, even when negative classes are present in the same image.

While we used horizontal concatenation to include the negative image in a single image, we believe that constructing datasets with images that naturally include both the target visual concept and its similar classes would further enhance the evaluation of personalized open-vocabulary semantic segmentation.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[3] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 1, 4

[4] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, 2023. 1, 4