USE: Universal Segment Embeddings for Open-Vocabulary Image Segmentation

Xiaoqi Wang[1,2,3] Wenbin He[1,2] Xiwei Xuan[1,2,4] Clint Sebastian[2] Jorge Piazentin Ono[1,2]
Xin Li[1,2] Sima Behpour[1,2] Thang Doan[1,2] Liang Gou[1,2] Han-Wei Shen[3] Liu Ren[1,2]
[1]Bosch Research North America [2]Bosch Center for Artificial Intelligence (BCAI)
[3]The Ohio State University [4]University of California Davis
wang.5502@osu.edu wenbin.he2@us.bosch.com xwxuan@ucdavis.edu, clint.sebastian@de.bosch.com
{jorge.piazentinono, xin.li9, sima.behpour, thang.doan, liang.gou}@us.bosch.com
shen.94@osu.edu liu.ren@us.bosch.com

A. Supplementary

In this supplementary, we include:

- visualization of curated data (A.1)
- additional ablation study (A.2)
- additional quantitative results on open-vocabulary semantic segmentation (A.3)
- qualitative comparison of open-vocabulary semantic segmentation methods (A.4)
- visualization of text-based query (A.5)

A.1. Examples of Curated Data

Figure 1 shows 4 example images with generated segment text pairs. The captions of the 4 images are listed as follows:

- Figure 1a:
  - A toy dinosaur standing on a sink next to a running faucet. (human-annotated)
  - a faucet running next to a dinosuar holding a toothbrush (human-annotated)
  - A toy lizard with a toothbrush in it's mouth standing next to a running water faucet in a bathroom. (human-annotated)
  - A fake toy dinousure has a green tooth brush in its mouth (human-annotated)
  - a sink with running water a mirror and a Godzilla toothbrush holder (human-annotated)
  - *In the picture, there is an orange dinosaur toy in front of a white sink. The dinosaur figurine is standing on its hind legs, with a blue toothbrush in its mouth. Above it is a mirror, and behind the dinosaur toy, there is a silver metal faucet. Next to the sink, there is a round hole. (MLLM-generated)*
- Figure 1b:
  - A picture of a dog laying on the ground. (human-annotated)
  - Dog snoozing by a bike on the edge of a cobblestone street (human-annotated)

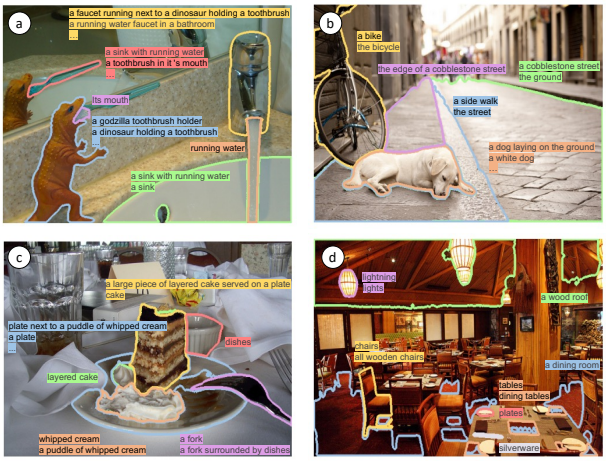

Figure 1. Examples of the curated segment text pairs.

- Figure 1c:
  - The white dog lays next to the bicycle on the sidewalk. (human-annotated)
  - a white dog is sleeping on a street and a bicycle (human-annotated)
  - A puppy rests on the street next to a bicycle. (human-annotated)
  - *A small dog is lying on the sidewalk, sleeping. There is a bicycle leaning against the wall with black spokes and a blue basket. On both sides of the street are tall buildings. In front of the building on the right, there are several people walking or standing. (MLLM-generated)*
- Figure 1c:
  - A slice of cake on a plate next to a puddle of whipped cream. (human-annotated)
  - A half-eaten piece of cake sits on a plate on a cluttered table. (human-annotated)
  - Multiple layer cake on plate, half eaten with a fork. (human-annotated)

Method	Type	Training Data	VL-Model	ADE-150	ADE-847	PC-59	PC-459	Average
USE	two-stage	COCO w/o GT mask	CLIP ViT-L/14	35.2	13.2	54.6	13.6	29.2
USE	two-stage	COCO	CLIP ViT-L/14	37.0	13.3	57.8	14.7	30.7

Table 1. Influence of combining models trained on different data sources. Performance is evaluated using mIoU.

Method	Training Data	Metric	ADE-150	ADE-847	PC-59	Average
SAN [2]	COCO	pAcc	71.0	57.2	78.8	69.0
USE	COCO	pAcc	70.7	58.7	77.4	68.9
SAN [2]	COCO	mAcc	50.7	25.3	77.5	51.1
USE	COCO	mAcc	58.4	29.8	77.3	55.1

Table 2. Open-vocabulary semantic segmentation benchmarks measured by pAcc and mAcc.

  - A nicely set table with a large piece of layered cake served on a plate. (human-annotated)
  - a piece of cake on a plate with a fork surrounded by dishes (human-annotated)
  - *The image shows a half-eaten, multi-layered cake in a large plate placed on a cluttered dining table. A fork is resting next to the plate, indicating that someone has already started enjoying their dessert. In addition to the cake and utensil, there are several other items scattered across the table: a bowl, multiple cups, two books, and some bottles. The scene appears quite busy with various objects occupying space on the table, suggesting it might be during a gathering or celebration where people have come together for food and drinks. (MLLM-generated)*
- Figure 1d:
  - A dining room filled with tables and chairs and lighting. (human-annotated)
  - A restaurant with tables, chairs, lights, plates and silverware. (human-annotated)
  - A restaurant has tables and chairs and a wood roof. (human-annotated)
  - A large very large restaurant with all wooden chair and tables. (human-annotated)
  - A room filled with dining tables and chairs. (human-annotated)
  - *This picture depicts a dining room with wooden floors and walls, and many tables and chairs arranged in an orderly manner. There are plate, fork, and spoon on each table, as well as napkins. In addition, there is a yellow lamp hanging above the table, and the ceiling is decorated with bamboo and wood structures. On the far right of the photo, there is a pillar with an ancient vase placed on it, and next to it is a wall covered with black curtains. (MLLM-generated)*

The MLLM-generated captions capture fine-grained details of the images. The generated segments with respect to the phrases in the captions are highly relevant, though there are still a few misaligned segments and phrases.

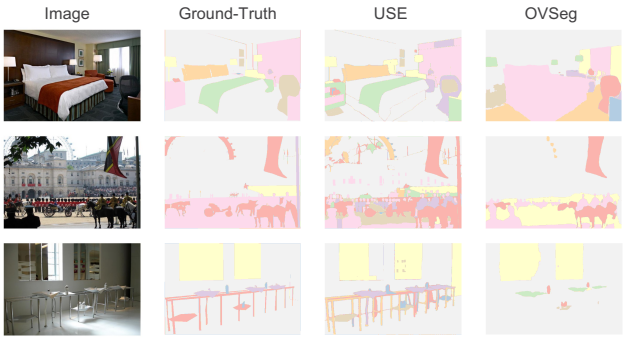

Figure 2. Visual results of open-vocabulary semantic segmentation on ADE20K validation set. Compared with OVSeg, our method predicts segmentation classes more accurately.

A.2. Ablation Study

Table 1 demonstrates the influence of combining models trained on different data sources, including human-annotated and curated masks. Same as OVSeg [1], different models' predictions are combined with different weights. We set the weight of the model trained on curated masks as 1. The weight of the model trained on human-annotated masks is set to 0.7 for the Pascal Context dataset with 59 categories and 0.25 for other datasets. We can see that by using human-annotated masks from the COCO dataset, the performance improves especially for the Pascal Context dataset with 59 categories. The reason is that the 59 categories of the Pascal Context dataset are similar to the human annotations in the COCO dataset [2]. For datasets that contain a large number of categories, the influence of human annotations decreases, for example, the ADE20K dataset with 847 categories.

A.3. Quantitative Results on Open-Vocabulary Semantic Segmentation

Table 2 shows additional quantitative results for open-vocabulary semantic segmentation on pixel accuracy (pAcc) and mean pixel accuracy (mAcc). Compared with SAN [2], our method archives comparable performance in terms of pAcc and obtains much higher mAcc, which indicates that our method performs better on small but challenging objects.

A.4. Visual Results on Open-Vocabulary Semantic Segmentation

Figure 2 shows visual results of open-vocabulary semantic segmentation on the ADE20K validation set with 150 categories. Compared with OVSeg, our segment classification results are more accurate and precise.

A.5. Visual Results on Text-based Query

Figure 3 shows the image segments retrieved from the ADE20K validation data for various phrases. We first obtain the segment embeddings for SAM-generated segments and the text embeddings of the phrases. Then we compute the similarity between the text and segment embeddings. We retrieve and visualize the top similar segments for each phrase. We observe that the segment embeddings can capture the semantic meaning of segments at various granularities, such as the fine-grained concept of "*car's license plate*". Meanwhile, the segment embeddings also capture the context information such as "*clear weather*", "*living room*", and "*the desk*".

References

[1] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted CLIP. In *CVPR*, pages 7061–7070, 2023. 2

[2] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 2, 3

Figure 3. Image segments retrieved for various phrases. The generated segment embeddings accurately capture the semantic meaning of the segments at various granularities.