CLIP-S^4: Language-Guided *S*elf-*S*upervised *S*emantic *S*egmentation

Wenbin He Suphanut Jamonnak Liang Gou Liu Ren
Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)
{wenbin.he2, suphanut.jamonnak, liang.gou, liu.ren}@us.bosch.com

A. Supplementary

In this supplementary, we include:

- the details of experimental setups and hyper-parameters (A.1)
- additional quantitative results on language-driven semantic segmentation (A.2)
- qualitative comparison of different language-driven semantic segmentation methods (A.3)
- visualization of pixel embeddings generated by different methods on Pascal Context (A.4)
- additional quantitative results on unsupervised semantic segmentation (A.5)
- visualization of the unknown classes learned from Pascal VOC 2012 (A.6)

A.1. Experimental Setup

Training During training, the augmented views of the training images are cropped into 336×336 using random cropping. The model weights are updated through SGD with momentum 0.9 and weight decay $1e^{-3}$. Our model is trained on one NVIDIA V100 GPU, which takes around 5 hours to train a model with 20k iterations.

Inference (Language-Driven) We use prompt-engineered texts with 85 prompt templates to generate text embeddings following [1, 6]. 80 prompt templates are the same as CLIP [4], and the rest 5 prompt templates are used for semantic segmentation tasks including:

- *there is a {} in the scene.*
- *there is the {} in the scene.*
- *this is a {} in the scene.*
- *this is the {} in the scene.*
- *this is one {} in the scene.*

To evaluate models' performance for *class-free semantic segmentation*, we split the 59 classes of Pascal Context into 4 folds as follows.

- Fold0: *airplane, bag, bed, bedclothes, bench, bicycle, bird, boat, book, bottle, building, bus, cabinet, car, cat*
- Fold1: *ceiling, chair, cloth, computer, cow, cup, curtain, dog, door, fence, floor, flower, food, grass, ground*
- Fold2: *horse, keyboard, light, motorbike, mountain, mouse, person, plate, platform, potted plant, road, rock, sheep, shelves, sidewalk*
- Fold3: *sign, sky, snow, sofa, table, track, train, tree, truck, tv monitor, wall, water, window, wood*

Inference (Unsupervised) We use k nearest neighbor (k-NN) search [3] and linear classification [5] to evaluate learned pixel embeddings. For k-NN search, each image is partitioned into 36 segments using k-means through 20 iterations. Each segment is assigned a class label based on the majority vote of the 20 nearest neighbors from the training set. For linear classification, we train an additional linear classifier, while fixing the pixel embeddings. The classifier is trained for 60k iterations. The learning rate starts at 0.1 and decayed with a polynomial learning rate policy. Training images are cropped into 512×512, and the batch size is set to 16.

A.2. Benchmarking Results on Language-Driven Semantic Segmentation

We show additional quantitative results for language-driven semantic segmentation on pixel accuracy (pAcc) and mean pixel accuracy (mAcc).

Tab. 1 shows the benchmarking results of MaskCLIP [6], MaskCLIP+ [6], and CLIP-S^4 on the COCO-Stuff validation set. Note that all class names are used during training for the sake of comparison. CLIP-S^4 consistently outperforms MaskCLIP and MaskCLIP+ on both metrics.

Tab. 2 shows the benchmarking results on Pascal Contex with unknown classes. The results are averaged over the 4 folds. CLIP-S^4 significantly outperforms MaskCLIP and MaskCLIP+ on both metrics.

Method	CLIP Model	COCO-Stuff pAcc	COCO-Stuff mAcc
MaskCLIP [6]	ResNet50	22.0	26.1
	ViT-B/16	27.7	32.4
MaskCLIP+ [6]	ResNet50	26.1	32.2
	ViT-B/16	35.7	39.9
CLIP-S^4	ResNet50	**28.6** (+2.5)	**36.2** (+4.0)
	ViT-B/16	**36.8** (+1.1)	**43.6** (+3.7)

Table 1. **Language-guided semantic segmentation benchmarks on COCO-Stuff with additional metrics (i.e., pAcc & mAcc).** CLIP-S^4 consistently outperforms the state-of-the-art methods on both metrics with CLIP models of different backbones.

Method	Pascal Context (w/ Unknown) pAcc	Pascal Context (w/ Unknown) mAcc
MaskCLIP [6]	45.2	46.4
MaskCLIP+ [6]	43.1±4.3	40.3±2.0
CLIP-S^4	**54.4±1.0**	**52.3±0.5**
vs. Baseline	**+9.2**	**+5.9**

Table 2. **Language-guided semantic segmentation benchmarks on additional metrics (i.e., pAcc & mAcc) for Pascal Contex with unknown classes.** The classes of Pascal Context are split into 4 folds with around 15 classes each fold. For each experiment, classes of one fold are considered as unknown. The results are averaged over the 4 folds. CLIP-S^4 significantly outperforms MaskCLIP and MaskCLIP+ on both metrics.

A.3. Visual Results on Language-Driven Semantic Segmentation

Fig. 1 shows some visual results of language-driven semantic segmentation on the COCO-Stuff validation set. The CLIP model with a ResNet50 backbone is used to train MaskCLIP+ and CLIP-S^4. Images are segmented with respect to different class names including both known and unknown classes. Here a known class means its prototype is used during training, and an unknown class means its name is not given to the model during training. Compared with MaskCLIP and MaskCLIP+, our results are more accurate for both known and unknown classes.

A.4. Visualization of Pixel Embedding Projections

Fig. 2 shows the projection of pixel embeddings generated by different methods on the Pascal Context dataset. For each class, we compute the average pixel embedding for different methods and project the embeddings of all classes using uniform manifold approximation (UMAP). We observe that the pixel embeddings generated by our method are well-aligned with CLIP for most of the classes. Meanwhile, MaskCLIP+ distorts the pixel embeddings with respect to the text embeddings.

Figure 1. **Visual results of language-driven semantic segmentation on COCO-Stuff validation set.** Our results are more accurate for both known and unknown classes compared with MaskCLIP and MaskCLIP+.

Method	pAcc	mAcc
SegSort [3]	86.9	-
ConceptContrast [2]	89.7	-
MaskCLIP [6]	91.9	77.6
MaskCLIP+ [6]	90.9	78.6
CLIP-S^4	**93.2** (+1.3)	**83.0** (+4.4)

Table 3. **Unsupervised semantic segmentation benchmarks on additional metrics (i.e., pAcc & mAcc).** CLIP-S^4 consistently outperforms the state-of-the-art methods on both metrics.

A.5. Benchmarking Results on Unsupervised Semantic Segmentation

We show additional quantitative results for unsupervised semantic segmentation on pixel accuracy (pAcc) and mean pixel accuracy (mAcc) for Pascal VOC 2012 (Tab. 3). The results are computed using k-NN approach [3]. Our method consistently outperforms the state-of-the-art unsupervised and language-guided approaches on both metrics.

Tab. 4 shows per-class IoU on Pascal VOC 2012 for unsupervised semantic segmentation. Our method outperforms the state-of-the-art unsupervised and language-guided approaches in most of the classes.

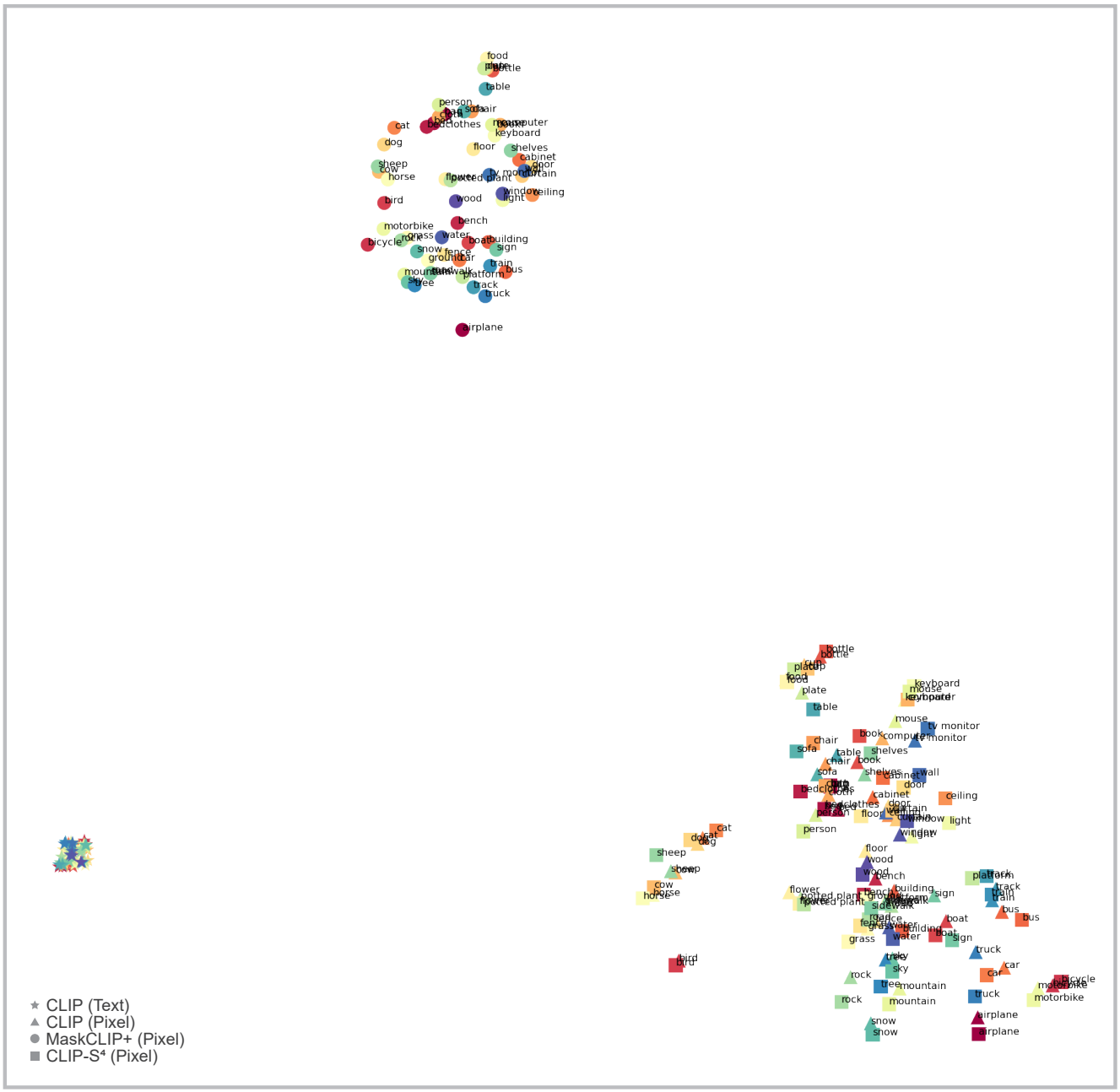

Figure 2. **Visualization of pixel embedding projections on Pascal Context for different methods.** For each class, the average pixel embedding is computed and projected with UMAP. Our method can generate pixel embeddings that are well aligned with CLIP for most of the classes, while MaskCLIP+ distorts the pixel embeddings of CLIP.

A.6. Unknown Classes Learned from Pascal VOC 2012

Fig. 3 visualizes examples of unknown classes learned from Pascal VOC 2012. We observe that various background classes such as sky, water, grass, and building can be extracted from the data without the guidance of class names. Meanwhile, we also find that some fine-grained classes can also be learned such as human hair and leg.

Fig. 4 visualizes the image segments extracted from Pascal VOC 2012 validation set. The image segments are placed based on the projection of the segment embeddings using UMAP. The color of the box around each image patch represents the class type of the segment, where green means known class and orange means unknown class. We observe that known classes form distinctive clusters such as the ta-

Method	Background	Aero	Bike	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Table	Dog	Horse	MBike	Person	Plant	Sheep	Sofa	Train	TV
SegSort	87.6	59.8	26.4	58.3	39.9	46.2	64.0	61.7	69.5	9.5	31.3	30.4	59.4	36.2	52.5	63.4	16.5	42.4	26.8	58.9	52.1
MaskContrast (Sup.)	-	76.2	26.9	70.2	49.6	56.1	80.3	66.8	66.8	10.6	55.1	17.5	65.2	51.8	59.7	58.8	23.1	73.5	24.9	70.9	38.9
ConceptContrast	89.5	71.0	30.9	73.6	53.2	60.4	80.5	73.9	78.6	17.4	47.9	47.7	68.5	51.6	63.7	71.8	36.3	57.7	31.7	72.6	55.4
HSG	-	75.1	32.2	76.9	60.4	63.9	81.7	75.5	82.0	18.5	48.7	51.2	71.5	55.0	69.4	71.0	39.8	66.8	33.3	72.3	59.6
MaskCLIP	90.5	69.7	**43.0**	78.4	59.4	55.5	78.3	77.7	83.5	**30.9**	77.4	48.3	80.5	77.1	73.1	79.1	**59.9**	77.9	39.5	77.6	56.2
MaskCLIP+	89.5	66.2	34.1	75.4	56.3	64.4	84.4	78.0	82.9	25.2	75.7	48.5	77.7	74.2	72.8	74.7	43.6	70.8	37.2	76.0	58.6
CLIP-S[4]	**92.4**	**83.5**	37.6	**85.7**	**69.4**	**75.3**	**89.5**	**79.9**	**87.9**	28.2	**83.1**	**57.5**	**83.9**	**82.4**	**76.7**	**80.1**	50.6	**80.5**	**42.4**	**81.4**	**63.2**

Table 4. **Per-class results on Pascal VOC 2012 for unsupervised semantic segmentation (mIoU).** Our method outperforms previous work consistently.

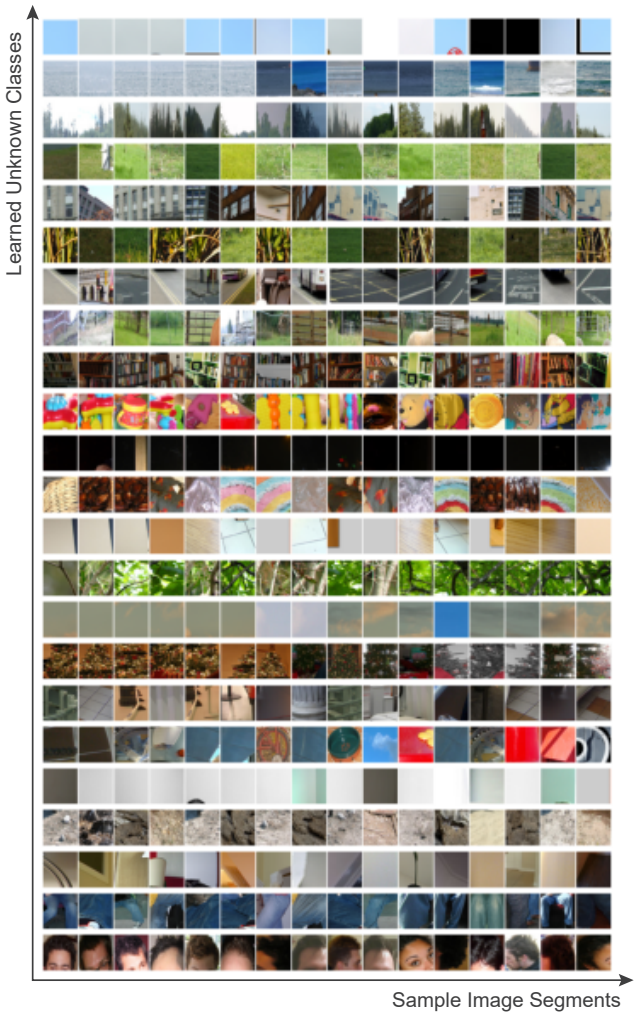

Figure 3. **Unknown classes learned from Pascal VOC 2012.** Various background classes (e.g., sky, water, grass, building, etc.) and fine-grained classes (e.g., human hair and leg) are learned from Pascal VOC 2012 without using class names.

ble (top left) and cow (top right). Between different known classes, unknown class clusters can be observed. Most of the unknown classes represent different types of backgrounds, such as grass (middle right), sky (bottom right), and water (bottom left). Some of the unknown classes represent sub-classes of known classes, such as different human body parts (middle left).

References

[1] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1

[2] Wenbin He, William Surmeier, Arvind Kumar Shekar, Liang Gou, and Liu Ren. Self-supervised semantic segmentation grounded in visual concepts. In *IJCAI*, pages 949–955, 2022. 2

[3] Jyh-Jing Hwang, Stella X. Yu, Jianbo Shi, Maxwell D. Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. SegSort: Segmentation by discriminative sorting of segments. In *ICCV*, pages 7333–7343, 2019. 1, 2

[4] 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*, pages 8748–8763, 2021. 1

[5] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, pages 10032–10042, 2021. 1

[6] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *ECCV*, 2022. 1, 2

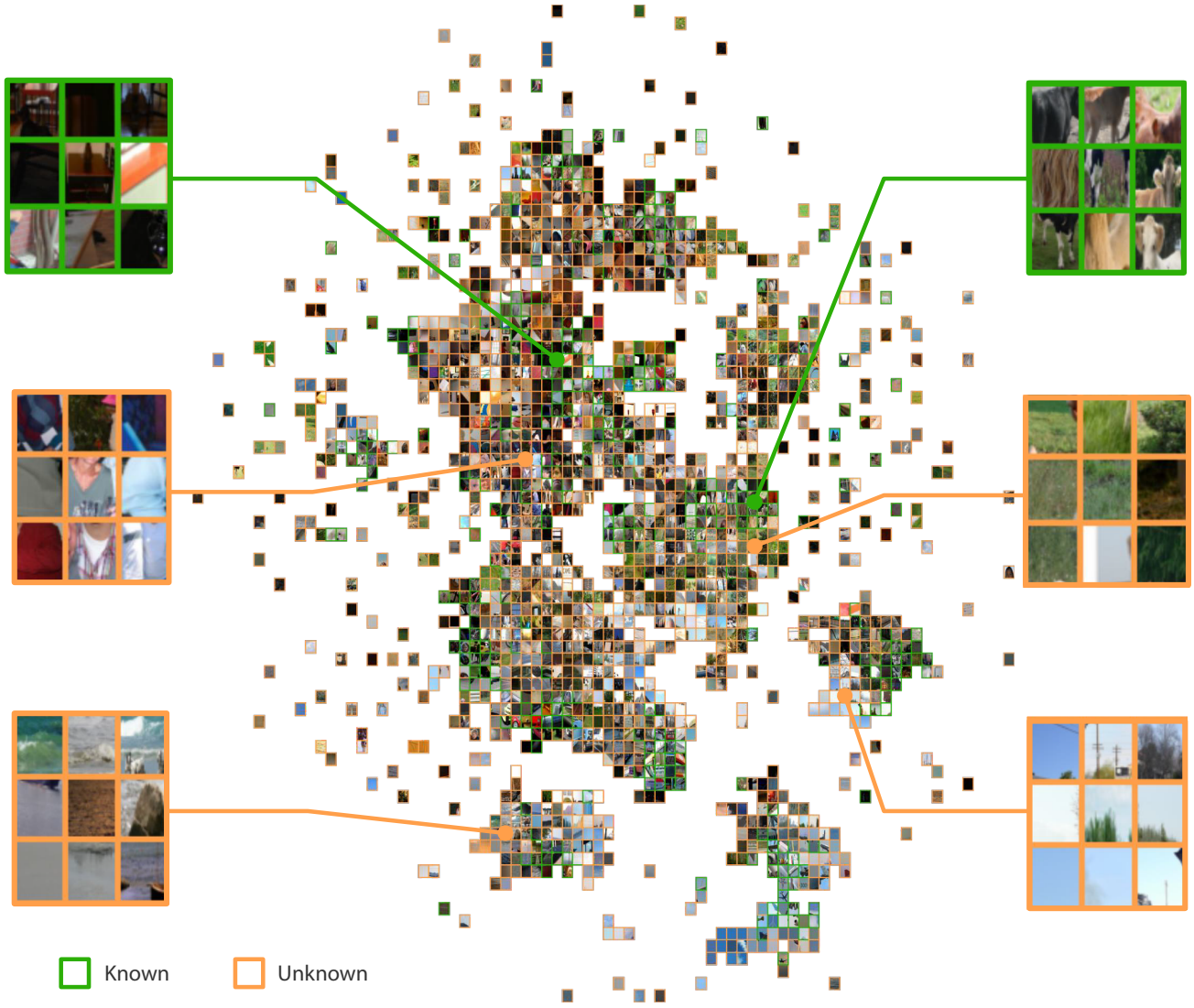

Figure 4. **Visualization of image segments from Pascal VOC 2012 validation set.** The image segments are placed based on the projection of the segment embeddings using UMAP. The color of the box around each image segment represents the class type (green for known classes and orange for unknown classes.)