

# Appendix of “ACSeg: Adaptive Conceptualization for Unsupervised Semantic Segmentation”

Kehan Li<sup>1,3</sup>   Zhennan Wang<sup>2</sup>   Zesen Cheng<sup>1,3</sup>   Runyi Yu<sup>1,3</sup>   Yian Zhao<sup>5</sup>   Guoli Song<sup>2</sup>  
Chang Liu<sup>4</sup>   Li Yuan<sup>1,2,3\*</sup>   Jie Chen<sup>1,2,3\*</sup>

<sup>1</sup> School of Electronic and Computer Engineering, Peking University, Shenzhen, China

<sup>2</sup> Peng Cheng Laboratory, Shenzhen, China

<sup>3</sup> AI for Science (AI4S)-Preferred Program, Peking University, Shenzhen, China

<sup>4</sup> Department of Automation and BNRist, Tsinghua University, Beijing, China

<sup>5</sup> Dalian University of Technology, Dalian, China

## A. Limitations

The proposed ACSeg directly exploits the pixel-level representations of a pre-trained ViT model. Although the ACG accurately groups pixels to concepts in the representation space, it can not be guaranteed that all representations reflect the corresponding semantic relationship unambiguously, especially when there is a gap between the pre-training dataset and the downstream task. On the other hand, the region-level representation is also transferred from pre-trained models and thus suffers from the domain shift. Although it performs well on VOC, the gap between the pre-training data of the backbone (ImageNet) and COCO causes only modest performance on COCO. It can be mitigated by training a task-specific model to produce better representations like STEGO [3] and SlotCon [18]. We take solving this issue as future research. Meanwhile, the pre-defined number of prototypes is necessary as the optimal partition of images is unknowable without given granularity. This hyperparameter impacts the granularity since each pixel pair is assigned to the closest prototype when optimizing the loss. Empirically, our method performs well when the variance of image complexity is not so large and this hyperparameter is determined by observation on several samples. Dealing with extremely large and complex datasets is still a limitation.

## B. Additional Implementation Details

**Dataset.** We use the PASCAL VOC 2012 [2] dataset (with extra augmentation data [4]) and COCO-Stuff [8] dataset for training and evaluation. For the COCO-Stuff dataset, we exploit the 27-classes subset and the “curated” split<sup>1</sup> introduced by IIC [6].

**Baseline.** Since the ACSeg can be regarded as a kind of clustering, we adopt some commonly used clustering methods k-means [5], spectral clustering [16], affinity propagation [1], and agglomerative clustering [10] as baselines for comparison. We use the implementation of these algorithms in Scikit-learn [11]. For these baselines, it is difficult to choose a fixed set of parameters for all images, which is why these methods cannot achieve good adaptiveness. We chose relatively suitable hyperparameters for different baselines, as shown in Table III. On the other hand, there are some existing over/ under-clustering methods for replacing the ACG, such as LOST [14] and DSM [9]. We compare the proposed ACG with them and show the results in Table I.

**K-means clustering.** In this setting, we run k-means clustering on the region-level representations produced by the concept classifier to get the class prediction of each concept. For the VOC 2012 dataset, we first recognize the concepts belonging to background as mentioned in Section 3.5. After that, we run k-means to assign the representations of predicted foreground concepts to 20 clusters and finally get predictions of 21 classes (20 foreground classes + 1 background class). The background

---

\*Corresponding author.

<sup>1</sup><https://www.robots.ox.ac.uk/~xuji/datasets/COCOStuff164kCurated.tar.gz>

class is recognized by the method proposed in Section 3.5. We show the results with some other possible alternatives in Table II. For the COCO-Stuff dataset, since there is no background category, we directly cluster all representations into 27 classes. The evaluation is done by matching the predicted clusters with the ground truth by Hungarian algorithm [7].

LOST [14]	DSM [9]	ACG ( <i>Ours</i> )
18.2	36.8	<b>47.1</b>

Table I. Results of other baselines.

Max Area [9]	Unsupervised Saliency [17]	Attention ( <i>Ours</i> )
39.1	46.0	<b>47.1</b>

Table II. Results of other background classification methods.

**$k$ -NN retrieval.** We adopt the weighted  $k$ -NN classifier in this setting. Specifically, the soft label of a concept is calculated by weighted averaging one-hot labels of  $k$  most similar concepts by their similarity, where we use the cosine distance between region embeddings as the similarity. Finally, the category with the highest score in the soft label is used as the classification result of a concept. We generate labels for concepts in the training set by the most overlapping ground truth region. The evaluation is done on the *val* set of VOC 2012 and COCO-Stuff. For the VOC 2012 dataset, we chose the *train* and *aug*<sup>2</sup> sets as the training set. For the COCO-Stuff dataset, we only report the results produced by using the first  $10k$  samples of the *train* set in the main text, because it is very time-consuming to get the results of baselines. We show the results of our method when using all samples of the *train* set in Table IV.

Algorithm	Hyperparameters
K-means	n_clusters = 5, init = ‘k-means++’
Spectral clustering	n_clusters = 5, n_components = 5
Affinity propagation	damping=0.5, preference = -2
Agglomerative clustering	distance_threshold = 0.65, linkage = ‘average’

Table III. Hyperparameters for different clustering baselines k-means, spectral clustering, affinity propagation, and agglomerative clustering.

Dataset	Method	K=1	K=5
COCO	K-means	29.9	33.1
	Spectral	28.5	31.3
	ACSeg (Ours)	30.4	34.0
	ACSeg†(Ours)	<b>33.8</b>	<b>37.7</b>

Table IV. Additional  $k$ -NN retrieval results. † indicates using all samples.

**Unsupervised semantic segmentation with text.** We first generate the text-based classifiers using the text encoder of CLIP [12] and the pre-defined categories. Following [12, 19], the words of categories are wrapped to sentences by templates and the classifier for a category is the average of the corresponding wrapped sentences. For the VOC 2012 dataset, we use the background classifier mentioned in Section 3.5 and only construct the text-based classifier for 20 foreground categories. For the COCO-Stuff dataset, we first classify concepts to all the things and stuff categories defined in COCO and then map them to 27 classes following ReCo [13]. For the visual representations, we first get the pixel-level representations following MaskCLIP [19] with CLIP-ViT-B/16 and then produce the region-level representation for each concept by averaging the pixels within it.

## C. Additional Qualitative Results

We show the additional t-SNE [15] visualization of the pixel-level representations and discovered concepts, clustering results, and retrieval results in Figure I, Figure II, and Figure III, respectively. In addition, the visualization of the training process can be found in [acseg\\_video.mp4](#) in the supplementary material.

<sup>2</sup>samples in SBD [4] but not in *train* and *val* sets



Figure I. Additional t-SNE visualization of the pixel-level representations (marked with dots) produced by self-supervised ViT and the corresponding concepts discovered by the ACG (marked with stars). We mark the concepts found by the ACG in different colors.



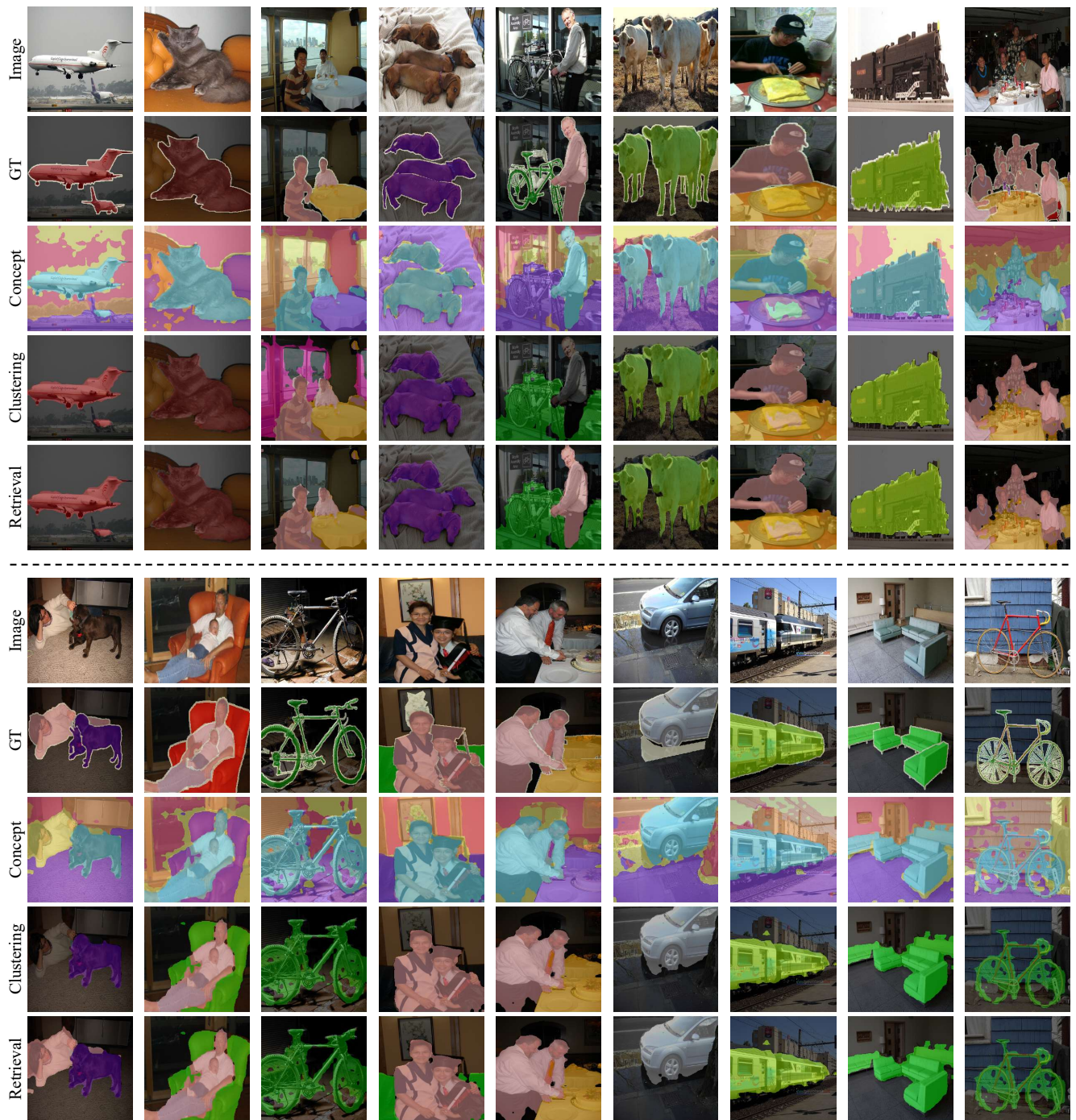


Figure II. Additional qualitative results on PASCAL VOC 2012 dataset.





Figure III. **Additional visualization of  $k$ -NN retrieval results.** We show five concepts with the highest similarity following each query concept (with red frame). The concepts is shown by the highlighted area in the image.

## References

- [1] Delbert Dueck. *Affinity propagation: clustering data by passing messages*. University of Toronto Toronto, ON, Canada, 2009. [1](#)
- [2] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007:1–45, 2012. [1](#)
- [3] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2021. [1](#)
- [4] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision*, pages 991–998. IEEE, 2011. [1](#), [2](#)
- [5] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975. [1](#)
- [6] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019. [1](#)
- [7] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. [2](#)
- [8] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014. [1](#)
- [9] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. [1](#), [2](#)
- [10] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012. [1](#)
- [11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. [1](#)
- [12] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. [2](#)
- [13] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. *arXiv preprint arXiv:2206.07045*, 2022. [2](#)
- [14] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *arXiv preprint arXiv:2109.14279*, 2021. [1](#), [2](#)
- [15] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. [2](#)
- [16] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007. [1](#)
- [17] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. [2](#)
- [18] Xin Wen, Bingchen Zhao, Anlin Zheng, Xiangyu Zhang, and Xiaojuan Qi. Self-supervised visual representation learning with semantic grouping. *arXiv preprint arXiv:2205.15288*, 2022. [1](#)
- [19] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision (ECCV)*, 2022. [2](#)