Open-Vocabulary Camouflaged Object Segmentation

Youwei Pang[1,2⋆], Xiaoqi Zhao[1,2⋆],
Jiaming Zuo[2], Lihe Zhang[1⋆⋆], and Huchuan Lu[1]

[1] Dalian University of Technology
[2] X3000 Inspection Co., Ltd
{lartpang,zxq}@mail.dlut.edu.cn
klaus@3000gy.com, {zhanglihe,lhchuan}@dlut.edu.cn

Abstract. Due to space constraints, some of the content from the main text has been arranged in this supplementary material.

1 More Model Details

1.1 Details of $\mathbf{E}_v$ and $\mathbf{T}_v$

In the proposed model, the visual encoder $\mathbf{E}_v$ and embedding layer $\mathbf{T}_v$ together are used to extract the high-level embedding corresponding to the object of interest in the input image. The two independent sub-networks are split from the visual network of CLIP [2, 10]. $\mathbf{E}_v$ contains all the feature encoding layers for extracting the multi-scale image features. And $\mathbf{T}_v$ corresponds to the final high-dimensional projection layer, which is used to convert the high-level image feature f^5 into the visual embedding vector f_v.

1.2 Multi-scale Image Features $\{f^i\}_{i=1}^5$

The multi-scale image features $\{f^i\}_{i=2}^5$ is from the four stages of ConvNeXt [8] with different output resolutions, respectively. While the feature f^1 is obtained by up-sampling the feature f^2 2× by bilinear interpolation.

1.3 Class Embedding

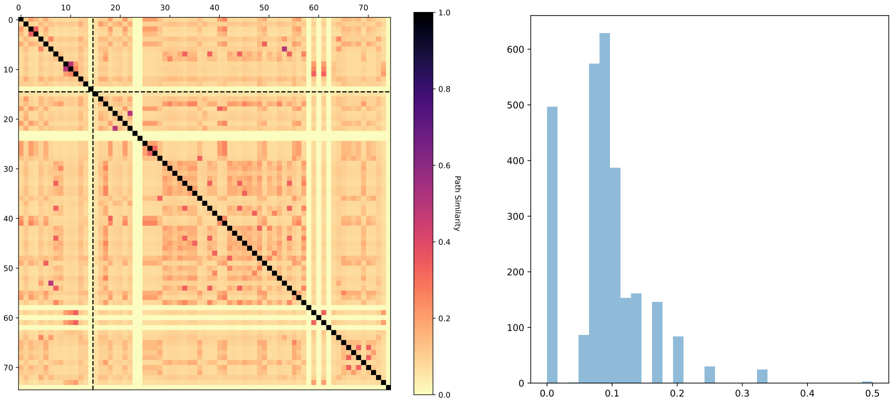
In the proposed algorithm, for each image input, the textual encoder needs to extract text embedding for all class texts. However, due to the nature of the algorithm design, these text embeddings are shared in each iteration, so they can be pre-computed in the inference to save inference cost.

⋆ Equal Contribution.
⋆⋆ Corresponding Author.

1.4 Details of Classification

During the inference, the class prediction P_c is generated from the pair-wise correlation matrix M_{cor} after the softmax operation, where M_{cor} is from the multiplication between the normalized visual embedding and textual embedding, *i.e.* f_v and f_t. Besides, in our experiments, introducing classification supervision on top of the existing form interferes with the training process of the model, which brings about significant performance degradation, with relative reductions of 8.6% and 15.7% for cS_m and cF_β^ω, respectively. Therefore, we do not consider classification loss during training, and M_{cor} is only used to construct the top-down iterative guidance.

2 More Ablation Studies



(a) Semantic similarity score map. **(b)** Semantic similarity score histogram.

Fig. 1: Class semantic similarity of OVCamo based on the Open English WordNet [9]. The classes belonging to the training and testing sets are separated here using black dashed lines in (a). Note that only the similarity between two different classes is considered in (b).

2.1 Class Semantic Similarity

To analyze the semantic similarity between the relabelled classes, we compute the path similarity between the classes based on the Open English WordNet [9] as shown in Fig. 1. Specifically, when p is the length of the shortest path between two classes, the path similarity score s is:

$$s = \frac{1}{p+1} \tag{1}$$

The score s ranges between 0.0 and 1.0, where the higher the score is, the more similar the two classes are. The score s is 1.0 when a class is compared to itself, and 0.0 when there is no path between the two classes (i.e., the path distance is infinite). As shown in Fig. 1b, the semantic similarity between classes in our class set $\mathcal{C}$ is very low. Most of them lie around 0.1, while the maximum is only 0.5. Such low similarity can better alleviate the complexity due to class semantic similarity during open vocabulary evaluation.

2.2 Class Hierarchy Relationships

In Fig. 2, we use the alluvial graph to show the class relationships at different levels, including super, base, and sub-classes. The sub-classes shown here include class names with clearer meanings preserved from the original data and class names after initial manual correction. The base classes represent the class names obtained after careful manual filtering and merging, which was used in all the experiments in this paper. The super classes generalize the base classes from a broader perspective.

3 Limitations and Future Works

3.1 Class Embedding Setting

In our proposed algorithm, the class embedding setting follows the existing OV-SIS methods. Although our iteration process does not impose much computational burden due to our caching mechanism, there is still a computational complexity associated with the number of classes. This is not ideal for practical applications in open vocabulary scenarios. More flexible and efficient class embedding designs are still worth exploring.

3.2 CLIP-based Architecture

The transfer application of the CLIP in downstream dense prediction tasks is limited by its pre-training form and the camouflage scenes that we focus on may be more affected. Some work [3,6,13] attempts to further finetune CLIP, but the finetuning strategy still needs to be designed more carefully due to the potential disruption to the CLIP's open vocabulary ability. And it also suggests that there is room for further improvement. Besides, as mentioned in the main text, the current ideal performance of the CLIP-based architecture is still far from the limit, which suggests that future breakthroughs in this field may require more powerful paradigms.

3.3 Data Scale

Although the number of finely labeled samples in our proposed dataset is over ten thousand, the data scale is still smaller than existing large datasets such as

COCO-Stuff [1] and ADE20K [18]. So there is still a need to collect more data, especially for classes with fewer samples. We can resort to rich web images, which also may lead to significant manual labeling costs. In addition, the data synthesis technique has been demonstrated in some recent work [15,16] to greatly facilitate the performance of semantic image segmentation tasks. This technique based on object masks and textual descriptions in existing datasets, may bring some new insights. But this also needs to address the ensuing interference problem.

3.4 Class Scale

Open-vocabulary segmentation as a hot topic, currently focuses on how to use the open-vocabulary capability of VLMs to segment objects with unseen classes. More data classes will indeed facilitate the development of OVCOS. However, the imperceptibility of camouflaged objects poses a great challenge for further expansion. This is the focus of our future work.

3.5 Prompt

The importance of prompt engineering for visual language modeling can be reflected in the existing literature [3, 5, 6, 10, 12–14, 17] and the experiments in this paper. The prompt forms used in this paper rely heavily on manual design, which is still limited by the knowledge of the prompter. More automated prompt-generation strategies may be needed in the future, which deserve more attention and exploration.

4 Dataset Copyright

We have investigated the copyright information of these data sources, and they are currently now widely used in the CSU field [4], and available for non-commercial academic use. Much of existing work [7,11] provides only new annotations and an index of the original data rather than the data itself. And users can download original data from the sources. Following these existing community practices, for the proposed OVCamo dataset, we list the data links provided by the original authors in the documentation. We also provide the class annotations we created and a detailed description of the way the data is organized in OVCamo. In addition, we thank the contributors of the relevant datasets in the our acknowledgements.

References

1. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2018) 4

2. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2022) 1
3. Cho, S., Shin, H., Hong, S., An, S., Lee, S., Arnab, A., Seo, P.H., Kim, S.W.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. arXiv preprint (2023) 3, 4
4. Fan, D.P., Ji, G.P., Xu, P., Cheng, M.M., Sakaridis, C., Van Gool, L.: Advances in deep concealed scene understanding. Visual Intelligence (2023) 4
5. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: International Conference on Learning Representations (2021) 4
6. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2022) 3, 4
7. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2016) 4
8. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2022) 1
9. McCrae, J.P., Rademaker, A., Fellbaum, C.D.: English wordnet 2019 – an open-source wordnet for english. In: Global WordNet Conference (2019) 2
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (2021) 1, 4
11. Unal, O., Dai, D., Gool, L.V.: Scribble-supervised lidar semantic segmentation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2022) 4
12. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2023) 4
13. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2023) 3, 4
14. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: Proceedings of European Conference on Computer Vision (2021) 4
15. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: Freemask: Synthetic images with dense annotations make stronger segmentation models. In: International Conference on Neural Information Processing Systems (2023) 4
16. Ye, H., Kuen, J., Liu, Q., Lin, Z., Price, B., Xu, D.: Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. ArXiv (2023) 4
17. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In: International Conference on Neural Information Processing Systems (2023) 4
18. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017) 4

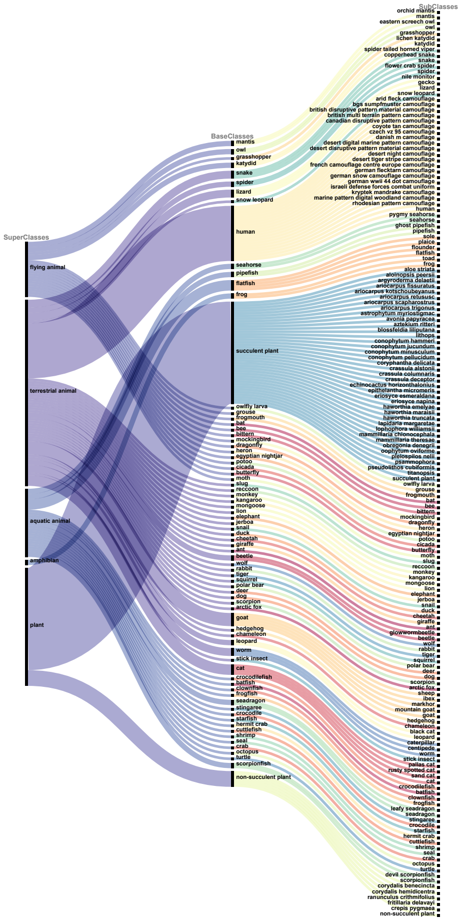


Fig. 2: Hierarchy of sample classes contained in the proposed OVCamo. **Only the base classes are used in our experiments.** Sub-classes that do not meet the criteria are simply removed and are not displayed here.