Exploring Open-Vocabulary Semantic Segmentation with CLIP

Xiangnan Fang
School of Computer Science
Renmin University of China
Beijing, China
2022201560@ruc.edu.cn

Abstract

The paper explores how to leverage CLIP for open-vocabulary semantic segmentation without additional training. We first replicate several existing models and experiment with replacing its vision backbone with a diffusion model to investigate potential performance gains. In addition, we introduce adjustments to the CLIP-generated features to enhance segmentation accuracy. To showcase our results, we developed a lightweight web application that enables users to upload images and visualize segmentation outcomes interactively. This project provides a hands-on opportunity to analyze the model's strengths and limitations, as well as to explore potential directions for improvement.

1 Introduction

Semantic segmentation is a fundamental task in computer vision, with the aim of assigning semantic labels to every pixel in an image. Open-vocabulary semantic segmentation, which seeks to segment novel or unseen categories beyond the training data, has gained increasing attention due to its potential to overcome the limitations of fixed-category segmentation methods. Recent advances in multimodal models like CLIP have shown promise in bridging the gap between vision and language, providing robust representations for open-vocabulary tasks.

Despite these advancements, it remains challenging to effectively utilize CLIP for open-vocabulary semantic segmentation. Existing models, such as ProxyCLIP, have made progress in this direction. Motivated by these works, this project explores how to leverage CLIP for open-vocabulary semantic segmentation with an emphasis on improving segmentation accuracy and practical application.

Specifically, we replicate several existing models and experiment with using a diffusion model as the vision backbone to enhance feature extraction and propose criteria for selecting the vision foundational models. Additionally, we try to adjust CLIP-generated features to better highlight foreground objects. To demonstrate our results, we developed a lightweight web application that allows users to upload images and visualize segmentation output interactively.

This paper is organized as follows. Section 2 introduces the related work, Section 3 details our attempts, Section 4 presents experimental results and analysis, and Section 5 concludes the study with insights and future directions.

2 Related Work

2.1 Open-Vocabulary Semantic Segmentation

Semantic segmentation, the task of assigning semantic labels to every pixel in an image, is a cornerstone of computer vision. Traditional approaches rely on supervised learning with annotated datasets, where the model is trained to recognize a fixed set of predefined categories. However, this reliance on labeled data limits the ability of these models to generalize to unseen categories, a critical requirement for real-world applications where new concepts and categories are constantly encountered.

Open-vocabulary learning addresses these limitations by leveraging visual-language alignment to bridge the gap between seen and unseen classes. This approach enables models to learn from auxiliary language data, such as image captions, which are easier and cheaper to obtain compared to traditional box or mask annotations. Moreover, captions often contain diverse vocabulary, including novel class names, attributes, and object actions, making them a rich source of supervision. By incorporating language data, open-vocabulary learning significantly enhances the scalability and generalization of segmentation models.(Wu et al., 2024)

2.2 CLIP in Vision Tasks

Contrastive Language-Image Pretraining (CLIP)(Radford et al., 2021) has revolutionized vision language integration by aligning images and text into a shared embedding space using a massive dataset of image-text pairs. This enables zero-shot generalization, making CLIP highly adaptable to tasks such as image classification, object detection, and semantic segmentation.

The strength of CLIP lies in its use of natural language as an extensible label space, allowing nuanced understanding of categories and attributes. In semantic segmentation, CLIP can associate text embeddings of novel categories with pixel-level features, enabling open-vocabulary tasks without extensive annotations.

Recent works have adapted CLIP for segmentation, such as MaskCLIP enhancing pixel-level alignment without retraining and OpenSeg leveraging grounding losses to associate textual descriptions with segmentation masks. However, CLIP's pretraining for image-level alignment limits its spatial precision, prompting methods like pseudo-masks and enhanced backbones to address this gap.

This work builds on CLIP's flexibility, exploring its adaptation for open-vocabulary semantic segmentation to advance annotation-efficient and scalable scene understanding.

2.3 ProxyCLIP

ProxyCLIP(Lan et al., 2024) is a relatively new model built on top of CLIP. Its core idea is to integrate CLIP's powerful semantic understanding capabilities with the strong local spatial feature consistency of visual foundation models (VFMs). By utilizing a Proxy Attention Module (PAM), it fuses the image features extracted from the VFM with the attention scores from the last layer of CLIP. This approach mitigates issues such as erroneous attention focus and noisy activations caused by CLIP's tendency toward dispersed global attention. Consequently, it enhances local feature consistency, addressing the problem of unclear boundaries in segmentation tasks and achieving significant improvements over the baseline.

This method of integrating two models is highly compelling. Therefore, this work primarily focuses on replicating ProxyCLIP, replacing the visual foundation model, and optimizing the features generated by CLIP.

3 Our Attempts

3.1 problem definition

Open vocabulary learning the training dataset is a collection of data-label pairs, where each pair consists of an input x_i and its associated label y_i. In the scene we consider, the label can also contain bounding boxes or masks. In addition, the training dataset incorporates vision-aware language vocabulary data, represented as l_i, which can be image-caption data or vision-aware class name embeddings.The training dataset can be defined as:

$$D_{\text{train}} = \{(x_1, y_1, l_1), (x_2, y_2, l_2), \ldots, (x_n, y_n, l_n)\},$$

where the x_i is the training data, y_i is the label from the base class set C_B, and l_i is the associated language data from a large vocabulary space C_L.

However, C_L is not strictly required to contain C_B or C_N, which means the language vocabulary may not cover all the classes in the vision data.

Similar to the training dataset, the evaluation dataset consists of data pairs. Notably, the labels in the evaluation data may include novel classes. The evaluation dataset can be defined as:

$$D_{\text{eval}} = \{(x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_m, y'_m)\}$$

During the evaluation, open vocabulary methods need to predict y'_i given x'_i in the realm $y'_i \in (C_B \cup C_N)$.

CLIP As mentioned earlier, the CLIP(Radford et al., 2021) model plays a significant role in the visual domain. It can be defined as:

$$\{(x_i, y_i)\}_{i=1}^{N},$$

where x_i represents an image, and y_i denotes the corresponding textual description associated with the image.

CLIP leverages contrastive learning to train two encoders $img(x; \theta_x)$ and $text(y; \theta_y)$, aiming to maximize the similarity between corresponding image-text pairs.

PAM The Proxy Attention Module (PAM) serves as a pivotal component in ProxyCLIP(Lan et al., 2024). It is designed to integrate visual representations $x \in R^{L_x \times D_x}$ from the VFM backbones and value embeddings$v \in R^{n \times L_v \times D_v}$ from CLIP's last attention layer. And adopts normalization and masking strategies to accommodate different visual foundation models.The process of the proxy attention module is formulated as follows:

$$A = \gamma \left(xx^T - \frac{\beta}{L_v^2} \sum_{i,j} [xx^T]_{ij} \right),$$

$$\mathcal{M}_{ij} = \begin{cases} 0, & A_{ij} \geq 0, \\ -\infty, & A_{ij} < 0, \end{cases}$$

$$\text{Attn}_p = \text{SoftMax}(A + \mathcal{M}).$$

3.2 new modules

unet as vfm ProxyCLIP utilizes visual foundation models such as DINO as the VFM backbone. After researching works on image segmentation(Xu et al., 2023), I discovered that diffusion models, represented by UNet, are also capable of encoding image features. Therefore, I attempted to use UNet as the VFM backbone.

In a standard UNet architecture, the network consists of a downsampling path, an upsampling path, and skip connections that link corresponding layers between the two paths. Let the downsampling features be denoted as $\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_k\}$, and the upsampling features as $\{\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_k\}$, where $k = L/2$, and L is the total number of layers. Skip connections combine the downsampling and upsampling features at each level via a concatenation operation $\oplus$, forming:

$$\mathbf{h}_i = \mathbf{g}_i \oplus \mathbf{f}_{k-i+1}, \quad i = 1, 2, \ldots, k.$$

The feature representation of the second-to-last layer of UNet, $\mathbf{f}_{L-1}$, is defined as the output of the second-to-last upsampling layer combined with the corresponding skip connection.

$\mathbf{f}_{L-1}$ incorporate semantic information from the upsampling path as well as spatial details preserved through the skip connections, making them suitable for downstream tasks.

adjustments to features inspired by the work of CLIP Surgery(Li et al., 2024), a consistency attention module was introduced and the model adjust the features from clip by computing the similarity by multiplying the CLS features with the text features. The resulting similarity was then summed along the text feature dimension to generate a redundancy matrix. By subtracting this redundancy matrix from the normal results, the model's ability to distinguish between background and foreground was enhanced.

Methods	Stuff	Object
MaskCLIP	12.51	-
MaskCLIP+	15.16	-
ProxyCLIP	26.53	37.54
ProxyCLIP(feat adjust)	-	22.89

Table 1: Performance comparison between MaskCLIP and ProxyCLIP.

4 Experiments

4.1 Main results

Baseline comparison Table 1 presents a comparison of baseline methods. MaskCLIP(Dong et al., 2023) is designed to enhance CLIP's dense pixel prediction capability by removing the embedding layers of q and k, and replacing the embedding layer of v and the final linear layer with 1x1 convolutional layers to focus on pixel-level feature extraction. MaskCLIP+ further improves performance by incorporating pseudo-labels and self-training.

On the COCO-Stuff dataset, MaskCLIP and MaskCLIP+ achieved mIoU scores of 12.51 and 15.16, respectively, while ProxyCLIP reached 26.53, nearly doubling the performance.

However, the results of ProxyCLIP with feature adjustments were suboptimal. This may be due to the fact that the features from the visual foundation model already emphasize the foreground objects for CLIP, making further adjustments unnecessary, or because this experiment did not identify an effective method for feature adjustment.

Qualitative results Fig.1 shows the qualitative comparison results before and after feature adjustment.

4.2 Ablation study

VFMs	Arch	Object
dino	B/16	36.25
dino	B/8	37.54
sam	-	23.75

Table 2: Impact of different vfms.

Impact of the backbones By experimenting with different architectures and types of visual foundation models, it can be verified that proxy attention performs well across various VFMs.

Figure 1: Qualitative results of feature adjustment.

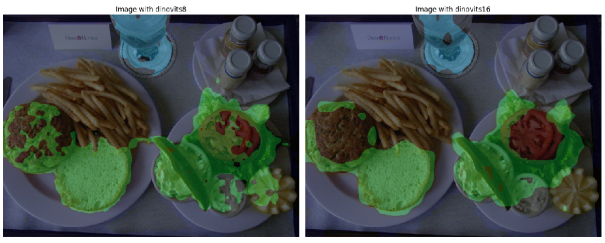

Figure 2: Qualitative results of different patchsize.

Qualitative results In Fig.2, the left side represents result from DINO S/8, while the right side is from DINO S/16. It can be observed that the segmentation on the left captures details more effectively. This indirectly supports the conclusion that the fine-grained features of diffusion models can provide guidance for CLIP's image segmentation.

Moreover, based on the experiments, a criterion for selecting VFMs can also be proposed. During the experiments where UNet was used to replace the VFM, it was observed that, since this approach does not require additional training, the training data of the pre-trained visual foundation model becomes particularly critical. A UNet model pre-trained on the Carvana dataset was used, but the results were suboptimal. This has been shown to be due to the limited and homogeneous nature of the training data.

5 Conclusion

lightweight web application We developed a lightweight web application with a simple front-end built using HTML and a back-end implemented with Flask. The application allows users to upload images, input the desired segmentation keywords, and obtain the segmentation results. Due to the use of tunneling techniques, the URL is not fixed and therefore not provided here.

conclusion In this paper, we explored non-training methods to enhance CLIP's image segmentation capabilities. Building on the concept of proxy attention, we conducted experiments on replacing VFMs and adjusting features. The results achieved demonstrate several key findings: first, the superiority of proxy attention was validated; second, a criterion for replacing VFMs was identified; and finally, a lightweight web application was developed.

Admittedly, there are several limitations to this study. For instance, we did not qualitatively investigate the impact of UNet on the results, nor did we evaluate the stability of the model. In the future, we aim to conduct more in-depth research to achieve better outcomes in this area.

References

Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. 2023. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *Preprint*, arXiv:2208.12262.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. 2024. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. *Preprint*, arXiv:2408.04883.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. 2024. A closer look at the explainability of contrastive language-image pre-training. *Preprint*, arXiv:2304.05653.

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. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, Bernard Ghanem, and Dacheng Tao. 2024. Towards open vocabulary learning: A survey. *Preprint*, arXiv:2306.15880.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. 2023. Open-vocabulary panoptic segmentation with text-to-image diffusion models. *Preprint*, arXiv:2303.04803.