

Training-Free Class Purification for Open-Vocabulary Semantic Segmentation

Qi Chen^{1,2} Lingxiao Yang¹ Yun Chen³ Nailong Zhao⁴ Jianhuang Lai¹
Jie Shao² Xiaohua Xie^{1*}

¹Sun Yat-sen University ²ByteDance Intelligent Creation ³University of Surrey ⁴Alibaba Cloud Computing
<https://github.com/chenqil126/FreeCP>

Abstract

Fine-tuning pre-trained vision-language models has emerged as a powerful approach for enhancing open-vocabulary semantic segmentation (OVSS). However, the substantial computational and resource demands associated with training on large datasets have prompted interest in training-free methods for OVSS. Existing training-free approaches primarily focus on modifying model architectures and generating prototypes to improve segmentation performance. However, they often neglect the challenges posed by class redundancy, where multiple categories are not present in the current test image, and visual-language ambiguity, where semantic similarities among categories create confusion in class activation. These issues can lead to suboptimal class activation maps and affinity-refined activation maps. Motivated by these observations, we propose *FreeCP*, a novel training-free class purification framework designed to address these challenges. *FreeCP* focuses on purifying semantic categories and rectifying errors caused by redundancy and ambiguity. The purified class representations are then leveraged to produce final segmentation predictions. We conduct extensive experiments across eight benchmarks to validate *FreeCP*'s effectiveness. Results demonstrate that *FreeCP*, as a plug-and-play module, significantly boosts segmentation performance when combined with other OVSS methods.

1. Introduction

Segmentation has achieved remarkable success with deep learning techniques [12, 14, 52], even in challenging semi-supervised [21, 63] and weakly-supervised [13, 43] settings. However, traditional segmentation models are limited to segmenting a small set of predefined classes within a closed vocabulary, which is much smaller than the number of classes used by humans to describe the real world. To address this, open-vocabulary semantic segmentation

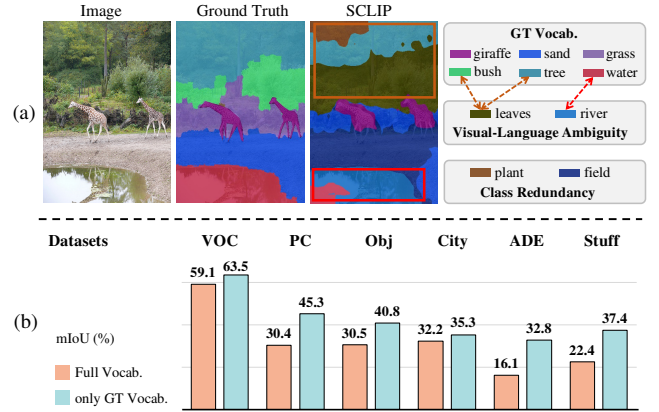


Figure 1. **Evidence** that overcomplete vocabulary affects OVSS performance. (a) Visualization of two types of problems: **Class Redundancy** and **Visual-Language Ambiguity**. (b) Performance comparison between full vocabulary and only GT vocabulary.

(OVSS) has been introduced to segment objects using arbitrary classes described by text.

Large-scale vision-language models (e.g. CLIP [38] and ALIGN [23]) have demonstrated impressive transferability in recognizing novel classes, which has recently been applied in OVSS [9, 35, 54]. The mainstream approach typically involves freezing the CLIP model while training newly added modules through mask supervision [22, 57] or text supervision [34, 54]. However, these methods require pixel-level annotated datasets (e.g. COCO Stuff [6]) or large-scale image-text datasets (e.g. CC-12M [10]), which demands significant computational resources. To address this limitation, recent approaches seek to exploit the localization capabilities of CLIP models to minimize training effort. These methods specify the forward pass for coarse localization, either by modifying the CLIP module [28, 32, 47, 66] or integrating pretrained vision foundation models [4, 26, 29, 48].

Although these methods localize objects without any training, they often yield sub-optimal results. The core issue lies in CLIP's design for image-text matching, which lacks the capacity for dense predictions. Therefore, CLIP

*Corresponding Author

often obtains imprecise segmentation and inaccurate classification. As shown in Fig. 1(a), one representative OVSS method – SCLIP [47] shows the significant discrepancies between the predicted and ground-truth classes. These discrepancies can be categorized into two main types: **Class (Text) Redundancy**, where predictions may include classes that are not actually present in current image (*e.g.*, ‘field’ and ‘plant’); and **Visual-Language Ambiguity**, which occurs when multiple semantically similar classes are present and these classes are strongly related to the same region. For instance, in the orange box in Fig. 1(a), this ambiguity appears among the classes ‘leaves’, ‘bush’, and ‘tree’, while in the red box, it occurs between the classes ‘river’ and ‘water’. To further investigate these issues, we conduct an analysis by restricting predictions to only ground-truth (GT) classes, as shown in Fig. 1(b). The results reveal substantial improvements in segmentation accuracy, particularly in scenarios with large vocabulary sets, which confirms that redundant ambiguous classes significantly impair model performance. These findings underscore the importance of mitigating class redundancy and visual-language ambiguity to enhance segmentation performance.

A naive approach is to first perform image-level multi-label classification and then use the recognized classes for segmentation. However, since classification occurs at the global level (image-level), it often misidentifies small or less prominent object classes as redundant and lacks the ability to detect semantically ambiguous classes in local regions. Alternatively, bottom-up methods first segment the image and then classify each segmented region [25, 41]. However, the performance of these approaches is highly dependent on the initial segmentation granularity, which can lead to fragmented results. In this paper, we aim to integrate classification and segmentation by leveraging classification cues to extract valuable local information, which is then utilized to refine class discrimination. On one hand, Class Activation Mapping (CAM) [64] can harness the CLIP model’s inherent discriminative classification capabilities for dense feature localization. Building upon this foundation, CLIP’s intrinsic self-attention mechanism can further refine the CAM-generated activation patterns, enhancing the model’s capacity to capture discriminative spatial-semantic features [33, 43]. On the other hand, the class-related spatial distribution cues provided by CAM and its refined version enable the identification of classes redundancy and visual-language ambiguity. As illustrated in Figure 2, it is achieved through quantitative analysis of spatial distribution differences between original CAM and refined CAM: 1) Class redundancy, where significant morphological variations emerge in activation regions of individual classes before and after refinement; 2) Visual-language ambiguity, manifested through substantial spatial overlap in activation patterns across multiple classes.

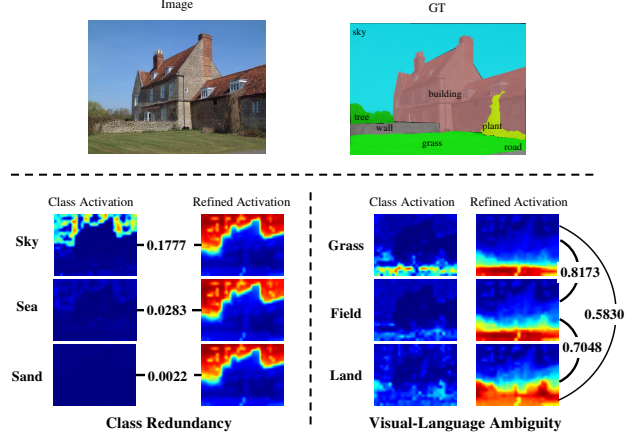


Figure 2. **Motivation of the proposed FreeCP method.** CAM and its refined version facilitate the identification of class redundancy and visual-language ambiguity by analyzing spatial distribution patterns. The reported values represent intersection-over-union (IoU), which quantify the consistency between activation maps, providing insights into the degree of similarity.

Motivated by these findings, we propose a training-**Free** Class Purification (FreeCP) method to enhance the performance of OVSS by purifying redundant and ambiguous classes. FreeCP begins by extracting image and text token features from CLIP, and then uses these features to compute image self-affinity and image-text affinity. The image-text affinity is then utilized to generate class-specific activation maps, which are further refined using image self-affinity. Subsequently, redundancy purification and ambiguity purification are applied in sequence to filter out redundant classes and resolve visual-language ambiguity. The remaining refined activations are finally used for segmentation prediction. Our approach achieves state-of-the-art performance on eight mainstream benchmarks, demonstrating the effectiveness of the proposed purification strategy. The main contributions of this paper are summarized as follows:

- We identify the problem of the class redundant and visual-language ambiguity caused by overcompleted vocabulary in OVSS, and provide an in-depth motivation by analyzing class activation maps.
- We propose a novel training-free framework, FreeCP, to purify classes. FreeCP first filters out redundant classes by examining spatial consistency between activations before and after refinement. Subsequently, it performs fine-grained recognition based on inter-class relationships to further solve visual-language ambiguous.
- Extensive experiments across eight benchmarks have showcased the state-of-the-art performance achieved by our FreeCP. And the generalization ability and effectiveness of class purification have been demonstrated.

2. Related Works

Existing OVSS methods can be categorized by their training approaches into four groups: fully-supervised, weakly-supervised, unsupervised, and training-free. Fully-supervised OVSS methods [16, 18, 19, 30, 56, 58, 60] initialize the model with pre-trained CLIP and then train it on large segmentation datasets. Weakly-supervised OVSS methods [7, 9, 11, 35, 39, 40, 51, 53–55, 59, 61] use image-text pairs as supervision, employing contrastive training to improve segmentation performance. Unsupervised methods [11, 50, 66] use self-supervised techniques to enhance CLIP’s dense prediction capabilities, avoiding the need for large image-text pair datasets. However, these methods still require substantial training computations, resulting in significant computational overhead.

Training-free methods require no additional training and have become a popular trend in the OVSS. MaskCLIP [66] modifies the self-attention layer of CLIP’s vision encoder by removing the self-attention pooling layer to produce pixel-level feature maps. CLIP Surgery [32], SCLIP [47], GEM [5], ProxyCLIP [29], and CLIPtrase [44] extend self-attention to more flexible and general formats, improving the segmentation ability. CLearCLIP [28] discards the residual connection and the feed-forward network of CLIP to achieve clearer and more accurate segmentation. Another line of work uses prototypes to leverage the robust correspondence in image representations for segmenting target objects. ReCo [45] employs CLIP to curate reference embeddings from unlabeled images, enhancing segmentation for rare concepts. Utilizing generative models, OVDiff [26] generates synthetic visual references from large text collections using diffusion models [42], and retrieves references based on input text for prototype-based segmentation. Additionally, FreeDA [4] visually localizes generated concepts, matching class-agnostic regions with semantic classes through local-global similarity. Further advancements include RIM [49], which integrates DINOv2 [37] and SAM [27] to construct well-aligned intra-modal reference features. CaR [46] introduces a recurrent framework that filters irrelevant text progressively. PnP-OVSS [24] combines text-to-image attention and salience dropout to iteratively acquire accurate segmentation of arbitrary classes. However, these methods often overlook redundancy and ambiguity among provided classes, which can lead to misclassification in localized areas.

3. Methods

In this section, we present our overall framework, FreeCP, as illustrated in Fig. 3. We begin by introducing fundamental formulation, CLIP and CAM in Sec. 3.1. Subsequently, we discuss the exploration of the activation refinement in Sec. 3.2. Finally, we provide a detailed explanation

of our training-free class purification method in Sec. 3.3.

3.1. Preliminary Background

Problem Formulation. Given an image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$ and a set of semantic words $\mathcal{Z} \in \mathbb{R}^K$, the objective of OVSS is to segment the pixels according to each word. Ideally, in OVSS, precise knowledge of the parameter K for each image or dataset is unavailable. Therefore, it is necessary to use a large K to encompass a sufficiently comprehensive range of categories. This ensures that the segmentation achieved is highly detailed and class-aware, facilitating fine-grained understanding. Previous approaches [19, 60] focus on improving models’ performances by pixel-level labels or image-text pairs. In this paper, we propose a simple yet effective **training-free method**, which can be easily plugged into multiple methods for OVSS.

CLIP-based Segmentor. CLIP [38] is trained to align images and texts globally. Thus, it has achieved remarkable progress in image-level understanding tasks such as classification [67], image-text matching [31], and image generation [62]. To adapt CLIP to pixel-level prediction tasks, prior work has explored the inherent properties of the CLIP image encoder, modifying it to support segmentation. The mainstream approaches [5, 32, 47] replace Q - K attention with V - V attention in the last block of the self-attention module. A few other approaches skip the last layer’s attention module entirely, taking V as the output [66]. Here, we adopt these modifications to enable more efficient segmentation. As shown in Fig. 3, the ViT-version of CLIP consists of an image encoder \mathcal{E}_I and a text encoder \mathcal{E}_T . The text encoder \mathcal{E}_T processes semantic words \mathcal{Z} with pre-defined prompts, *e.g.*, A photo of a $\{\mathcal{Z}\}$, to extract their text representations $\mathbf{T} = \mathcal{E}_T(\mathcal{Z}) \in \mathbb{R}^{K \times d}$, where d is the embedding dimension. The image encoder \mathcal{E}_I processes N patches of a single image to obtain N patch tokens $\mathbf{F}^p \in \mathbb{R}^{N \times d}$ and a class token $\mathbf{F}^c \in \mathbb{R}^{1 \times d}$. Inspired by CAM [64], we treat the text embeddings \mathbf{T} as class weights and the patch tokens \mathbf{F}^p as image features to compute the class-wise activation map $\mathbf{M} \in \mathbb{R}^{K \times \frac{H}{P} \times \frac{W}{P}}$, where P is the patch size and \mathbf{M} is defined as follow:

$$\mathbf{M}_j = \text{Reshape}\left(\frac{\exp(\text{Sim}(\mathbf{F}^p, \mathbf{T}_j))}{\sum_j \exp(\text{Sim}(\mathbf{F}^p, \mathbf{T}_j))}\right), \quad (1)$$

where $j \in [1, 2, \dots, K]$ is the index of the classes. $\text{Sim}(\cdot)$ denotes cosine similarity. $\text{Reshape}(\cdot)$ recovers the spatial structure of image patches by converting the 1-D patch embeddings back to 2-D maps.

3.2. Analysis: Explore the potential of refinement

Although CLIP combined with CAM exhibits certain class-aware localization capabilities, the quality of its dense prediction remains limited due to the absence of dense su-

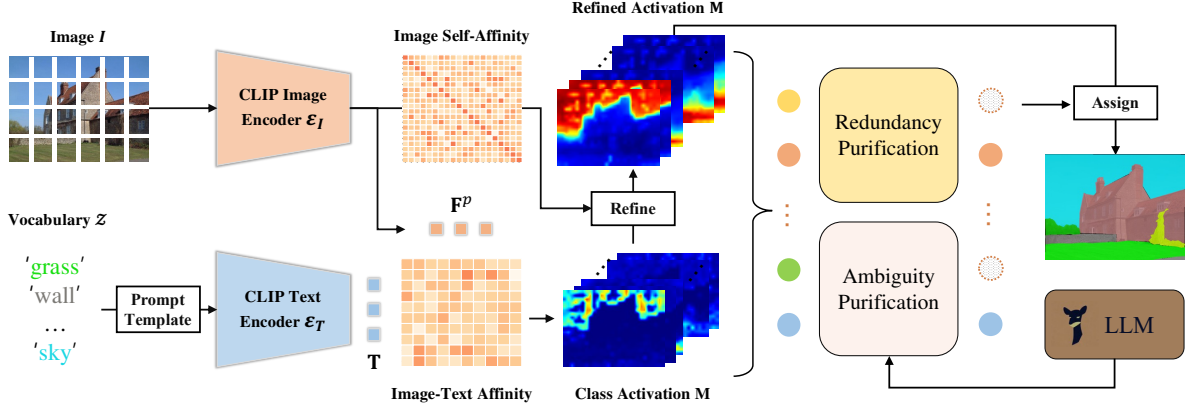


Figure 3. **Overview of the proposed FreeCP method.** Based on a ViT version of CLIP model, the methodology includes three sequential stages: **First**, image self-affinity and image-text affinity are derived through the image encoder \mathcal{E}_I and text encoder \mathcal{E}_T of CLIP. We then leverage image-text affinity to generate class-specific activations \mathbf{M} and refined activations $\tilde{\mathbf{M}}$ with image self-affinity. **Subsequently**, we formulate spatial consistency and perform redundancy purification and ambiguity purification to eliminate redundant and visual-language ambiguous classes. **Finally**, the retained class activations play a pivotal role in determining the conclusive segmentation prediction.

pervision during CLIP’s pre-training. Motivated by recent advances in weakly-supervised semantic segmentation [2, 33, 43], which leverage affinity matrices to enhance activation maps, we investigate the efficacy of refining CAM through a learned affinity matrix. By propagating activations across high-affinity patches, the resulting activation maps become more comprehensive and better capture the spatial extent of object regions. In our work, we adopt the self-attention matrices of the CLIP image encoder as the affinity matrix, capitalizing on their inherent ability to capture semantic relationships among image patches. Specifically, the image self-affinity matrix $SA \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times \frac{H}{16} \times \frac{W}{16}}$ is computed as the average of the resized self-attention matrices from multiple layers:

$$SA = \frac{1}{L} \sum_l \psi(A_l), \quad (2)$$

where A_l denotes the self-attention matrix at the l -th layer of the image encoder \mathcal{E}_I , L is the total number of layers considered, and $\psi(\cdot)$ is a bilinear interpolation resize operation applied to standardize the spatial dimensions of attention maps across layers. Subsequently, the affinity-based refinement is applied to the initial activation maps $\mathbf{M} \in \mathbb{R}^{K \times \frac{H}{16} \times \frac{W}{16}}$ as follows:

$$\tilde{\mathbf{M}}_i = \mathbf{M}_i \times SA, \quad (3)$$

where $i \in [1, 2, \dots, K]$ indexes the semantic classes.

However, in open-vocabulary settings, the set of semantic categories present in an image is not known a priori, which significantly impacts the behavior of affinity-based refinement. As shown in Tab. 1, when ground-truth class labels are available, the refinement process substantially improves the quality of the activation maps, as it can focus

Table 1. Comparison of using refined activation of GT vocabulary and that of full vocabulary on different datasets.

Methods	VOC	PC	Obj	City	ADE	Stuff
Baseline	59.4	29.7	33.4	32.0	15.6	22.1
Refine w/o GT	24.0	19.5	11.3	23.3	7.6	13.4
Refine w/ GT	72.0	50.7	42.9	35.6	37.5	42.5

exclusively on the relevant classes, leading to more complete and accurate localization. In the absence of ground-truth labels, the refinement results in a noticeable performance drop compared to the baseline. The degradation can be primarily attributed to the class-agnostic nature of the refinement mechanism, which unintentionally reinforces activation for irrelevant categories. This issue is visually exemplified in Fig. 2, where refinement introduces spurious activations in regions unrelated to any ground-truth class, thereby undermining overall segmentation quality.

To address the challenges posed by affinity-based refinement in open-vocabulary settings, we analyze its behavior from the structural patterns of activation consistency across CAMs. Specifically, we observe that: *If a class shows strong consistency in its activation map before and after refinement, it is more likely to be a true positive—that is, the class is genuinely present in the image. Conversely, if a class’s response changes significantly after refinement, this often indicates a redundant or irrelevant prediction.* Furthermore, we extend this observation to inter-class relationships: *When two or more classes exhibit highly similar refined activation maps, it indicates the presence of semantic or visual-language ambiguity between them.* These findings suggest that the consistency patterns it induces can serve as useful signals for post-hoc refinement control.

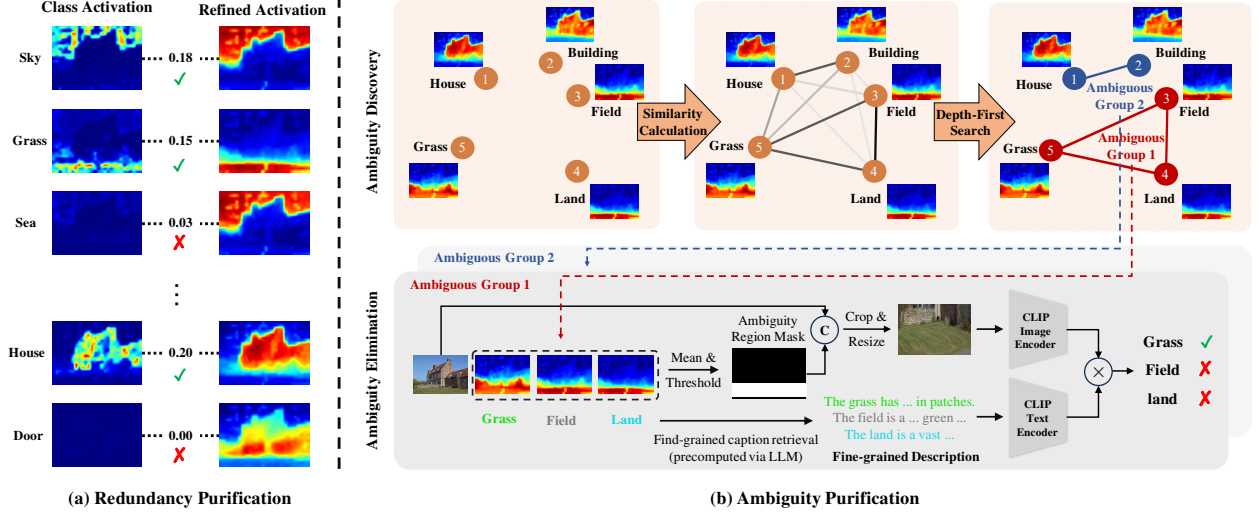


Figure 4. **Overview of Class Purification.** For all classes in the vocabulary, we first conduct (a) **Redundancy Purification**, eliminating classes whose intra-class spatial consistency falls below the predefined threshold. Subsequently, the remaining classes undergo (b) **Ambiguity Purification**, where visual-language ambiguity groups are identified based on inter-class spatial consistency. Fine-grained descriptions are then incorporated to resolve ambiguities within local regions.

3.3. Training-Free Class Purification

Inspired by aforementioned observations, we propose to use spatial consistency between these activation maps to solve the class redundancy and visual-language ambiguity. In this paper, we introduce a simple course-level metric – IoU to represent the Spatial Consistency (SC) between activation maps, which is defined as follows:

$$SC(\mathbf{X}, \mathbf{Y}) = \frac{\sum[\mathbf{X} \cdot \mathbf{Y}]}{\sum[\mathbf{X} + \mathbf{Y} - \mathbf{X} \cdot \mathbf{Y}]} \quad (4)$$

where \mathbf{X} , \mathbf{Y} are the activation maps, and (\cdot) represents element-wise multiplication.

In the following, we will present our training-free two-stage class purification: **Redundancy Purification** and **Ambiguity Purification**. Given that only a small number of classes are present in an image usually, we first identify and purify these redundant classes. Subsequently, among the remaining classes, we detect class groups exhibiting visual-language ambiguity and iteratively resolve the ambiguity using fine-grained descriptions.

3.3.1. Redundancy Purification

When comparing the class activation map with the refined activation map for the same class, differences in consistency appear between classes present in the image and redundant classes. As shown in the left side of Fig. 4 (a), the activation maps for ‘sky’, ‘grass’ and ‘house’ exhibit distinct response values, while other classes, such as ‘sea’ and ‘door’, display more subtle responses. We observe that the attention-based refinement contributes to enhancing the accuracy of objects

in the actual classes. Changing before and after refinement focus on more fine details. However, for classes like ‘sea’ and ‘door’, the refined activation maps introduce a significant number of irrelevant erroneous responses. Therefore, existing and redundant classes can be evaluated by analyzing the intra-class SC before and after refinement:

$$S_i = SC(\mathbf{M}_i, \tilde{\mathbf{M}}_i). \quad (5)$$

We empirically remove the i -th class if S_i falls below a predefined threshold T_{rp} . Consequently, the updated class set K' is obtained, where $K' \subseteq K$.

3.3.2. Ambiguity Purification

After filtering redundant classes, we then address the issue of visual-language ambiguity among the remaining classes. As shown in Fig. 4 (b), we propose two steps to solve the ambiguity purification: discovery and elimination.

Ambiguity Discovery: To identify visual-language ambiguity, we take the inter-class spatial consistency as an ambiguity indicator. Specifically, we calculate the SC between class i and class j for all possible pairs, which can be derived from Eq. (4) and shown as follows:

$$P_{i,j} = SC(\tilde{\mathbf{M}}_i, \tilde{\mathbf{M}}_j). \quad (6)$$

A higher $P_{i,j}$ indicates a greater likelihood of ambiguity between the two classes. Therefore, we set a threshold T_{ap} to binarize this probability, highlighting potential ambiguous classes, which is defined as follows:

$$P_{i,j} = \begin{cases} 1, & \text{if } P_{i,j} > T_{ap}, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

Subsequently, we can extract connected class groups using a Depth-First Search algorithm based on P . Each group consists of two or more mutually ambiguous classes and is defined as an ambiguity group. As shown in Fig. 4 (b), five classes are divided into two distinct ambiguity groups.

Ambiguity Elimination: For each discovered ambiguity group, we locally resolve the visual-language ambiguity with the help of fine-grained class description. First, we need to localize the ambiguous region. We average activation maps of all ambiguous classes to highlight high-response regions, which likely indicate ambiguity. Bounding boxes of high-response regions are extracted following [33]. And then the ambiguous regions are cropped from the original image and resized to a specified shape (e.g., 112×112), forming the cropped image $\hat{\mathcal{I}}$. With the cropped image $\hat{\mathcal{I}}$, we input it into the CLIP Image encoder to extract the visual feature of the ambiguous regions $\hat{\mathbf{F}}^c = \mathcal{E}_I(\hat{\mathcal{I}})$.

To achieve a more precise distinction among ambiguous classes, we employ a Large Language Model (LLM) to generate detailed and fine-grained text descriptions $\hat{\mathbf{T}}_k$ for each ambiguous class k . It is worth noting that for computational efficiency, we precompute and store fine-grained text descriptions for all classes within the dataset vocabulary. During inference, the corresponding text descriptions are retrieved directly, eliminating additional computational overhead. Ultimately, the final class of these ambiguous regions is determined by comparing the similarity between visual and textual features, as follows:

$$k^* = \arg \max_k \text{Sim}(\hat{\mathbf{F}}^c, \hat{\mathbf{T}}_k). \quad (8)$$

It is important to emphasize that no classes are removed in this step. Instead, classes that fail the competition are set to zero in the local region to avoid interference between verification results in different areas.

After processing all ambiguity groups, we apply the argmax operation to the refined activations of the remaining K' classes to determine the final segmentation results.

4. Experiments

4.1. Dataset and Evaluation Metric

Similar to previous works [47], we conduct experiments on five commonly used segmentation benchmarks: PASCAL VOC [20], PASCAL-Context [36], MS COCO [6], ADE20K [65], and Cityscapes [17]. **PASCAL VOC:** This object-centric semantic segmentation dataset contains 20 object classes and 1 background class. There are two variants of VOC: VOC21, which includes all 21 classes, and VOC20, which removes the background class to form 20 classes. **PASCAL-Context:** This dataset contains 5,105 validation images, with 459 classes. The most frequent 59 classes are used to form the PC59 version for evaluation, while we also evaluate the PC60 by treating all other

classes as background. **MS COCO:** This dataset comprises 5,000 densely annotated validation images, including 80 thing classes and 91 stuff classes, collectively forming the Stuff dataset. Object merges all stuff classes into a single background class, resulting in a total of 81 classes. **ADE20K:** This scene-parsing dataset includes 150 fine-grained classes. We evaluate 2,000 validation images, with an average of 9.9 classes per image. **Cityscapes:** It is designed for urban scene understanding, including 500 validation images with 19 semantic classes. **Evaluation metric:** For all experiments, we use mean Intersection-over-Union (mIoU) as the evaluation metric.

4.2. Implementation details

We adopt the ViT-B/16-based pre-trained CLIP as our default backbone. For generating fine-grained descriptions, we utilize the Vicuna-13b-1.5 [15]. We design three instructions to prompt LLM, and generate five answers for each instruction, for a total of 15 descriptions for each class. The final textual feature is the average of these 15 descriptions. The prompts and samples of generated descriptions are presented in supplementary materials. In inference, all images are resized such that the shorter side is 448 pixels (560 for Cityscapes), and a sliding window approach with a size of 384 pixels and a stride of 112 pixels is employed. The thresholds (T_{rp} , T_{ap}) are configured based on prior knowledge regarding the dataset’s semantic complexity and category granularity. For complex datasets like ADE and Stuff, which include 150+ fine-grained categories and dense scenes, a low threshold helps retain small or rare regions. Cityscapes and Context have moderate category counts and structured layouts, so we use a balanced threshold to maintain both precision and recall. Simpler datasets like VOC and Object, with fewer object-centric classes and clear foregrounds, benefit from a higher threshold to suppress noise. Our results do not involve any post-processing methods such as PAMR [3] or denseCRF. All evaluations are conducted using $8 \times$ NVIDIA RTX 3090 GPUs.

4.3. Comparison with state-of-the-art methods

Compared baselines. Following [28, 47], we compare our FreeCP with state-of-the-art OVSS methods across eight benchmarks, including three that consider background class and five that do not. The compared methods are classified into two groups. The first group consists of methods that rely solely on CLIP [38], including MaskCLIP [66], ReCo [45], SCLIP [47], CaR [46], GEM [5], CLIP-trase [44], OVDiff [26], and ClearCLIP [28]. The second group comprises methods that incorporate additional vision foundation models, including ZeroGuideSeg [41], RIM [48], FreeDA [4], LaVG [25], and ProxyCLIP [29].

Quantitative Results. As a strategic plug-in approach, our FreeCP can be integrated into existing training-free meth-

Table 2. **Comparison with the state-of-the-art training-free methods on eight benchmarks.** Our FreeCP can be integrated with existing training-free methods, leading to significant performance improvements on all benchmarks. Result with * is postprocessed with denseCRF.

Methods	With background			Without background					Avg.
	VOC21	PC60	Object	VOC20	City	PC59	ADE	Stuff	
<i>With additional models, e.g., DINOv2, Stable Diffusion, SAM</i>									
ZeroGuideSeg [ICCV'23] [41]	-	-	-	20.1	-	19.6	-	-	-
RIM [CVPR'24] [48]	77.8	34.3	44.5	-	-	-	-	-	-
FreeDA [CVPR'24] [4]	55.4	38.3	37.4	85.6	36.7	43.1	22.4	27.8	43.3
LaVG [ECCV'24] [25]	62.1	31.6	34.2	82.5	-	34.7	15.8	23.2	-
ProxyCLIP [ECCV'24] [29]	61.3	35.3	37.5	80.3	38.1	39.1	20.2	26.5	42.3
<i>Without additional models</i>									
CLIP [ICML'21] [38]	18.8	9.9	8.1	49.4	6.5	11.1	3.1	5.7	14.1
ReCo [ECCV'22] [45]	25.1	19.9	15.7	57.7	21.6	22.3	11.2	14.8	23.5
CaR* [CVPR'24] [46]	67.6	30.5	36.6	91.4	-	39.5	17.7	-	-
CLIPtrase [ECCV'24] [44]	53.0	30.8	44.8	81.2	-	34.9	17.0	24.1	-
OVDiff [ECCV'24] [26]	62.8	28.6	34.6	80.9	23.4	32.9	14.1	20.3	-
MaskCLIP [ECCV'22] [66]	43.4	23.2	20.6	74.9	24.9	26.4	11.9	16.7	30.3
MaskCLIP + FreeCP [Ours]	64.4	34.7	36.2	84.1	32.5	36.6	17.6	23.3	41.2
GEM [CVPR'24] [5]	56.9	32.6	31.1	79.9	30.8	35.9	15.7	23.7	38.3
GEM + FreeCP [Ours]	64.7	35.5	36.9	80.6	35.7	39.1	17.8	25.8	42.0
ClearCLIP [ECCV'24] [28]	51.8	32.6	33.0	80.9	30.0	35.9	16.7	23.9	38.1
ClearCLIP + FreeCP [Ours]	64.5	35.7	36.9	81.5	34.4	39.3	18.9	26.1	42.2
SCLIP [ECCV'24] [47]	59.1	30.4	30.5	80.4	32.2	34.2	16.1	22.4	38.2
SCLIP + FreeCP [Ours]	65.8	35.3	37.2	84.3	33.3	38.0	18.4	24.9	42.1

ods to reduce the impact of class redundancy and visual-language ambiguity on performance. As shown in Tab. 2, after applying FreeCP to MaskCLIP, GEM, ClearCLIP, and SCLIP, the mIoU of the original methods improved by 10.9%, 3.7%, 4.1%, and 3.9%, respectively. This improvement is attributed to the combined effect of the refinement techniques and class purification, which optimize object contours and enhance segmentation quality. It is worth noting that although MaskCLIP initially exhibited suboptimal performance, its performance improved significantly with FreeCP. This demonstrates that our method is robust to initial performance. Furthermore, FreeCP achieves comparable results on several datasets compared to methods that introduce additional models. All these results strongly validate the generalizability and effectiveness of our FreeCP.

4.4. Ablation studies and Analyses

In our ablation studies, we choose SCLIP [47] as the baseline to conduct extensive experiments, demonstrating the contribution and effectiveness of the proposed method.

Effect of class purification. We conduct an ablation study to evaluate the effectiveness of the core contribution of our FreeCP: redundancy purification (RP) and ambiguity purification (AP). The comparison results are presented in Tab. 3. As previously discussed, refining baseline prediction without true classes leads to a sharp decline in performance.

Table 3. **Ablations on class purifications.** The results of four purification options are presented. RP is Redundancy Purification, AP is Ambiguity Purification, and FreeCP uses the RP-AP option.

Methods	VOC21	PC60	Object	City	ADE	Stuff
Baseline	59.8	31.6	34.5	32.0	17.2	23.2
+ Refine	27.5	21.1	11.9	26.0	9.1	14.3
+ Purification options						
RP	65.8	35.1	37.2	33.2	17.8	24.1
AP	37.7	26.1	13.6	24.0	10.8	15.0
AP-RP	57.3	33.8	36.7	32.6	16.8	23.8
RP-AP	65.8	35.3	37.2	33.3	18.4	24.9

By adding redundancy purification, we remove a significant number of weakly-responsive incorrect classes, leading to a notable improvement over the baseline. Additionally, with the incorporation of ambiguity purification, our method achieves further improvements. Since class confusion is more pronounced in datasets with complex classes, the enhancement brought by ambiguity purification is relatively smaller compared to redundancy purification. When only AP is applied, the performance significantly declines due to the persistent interference of redundant classes. Even if RP is performed after AP, it merely removes redundant classes without compensating for errors introduced earlier. Fig. 5 visualizes examples from COCO Stuff and ADE20K.

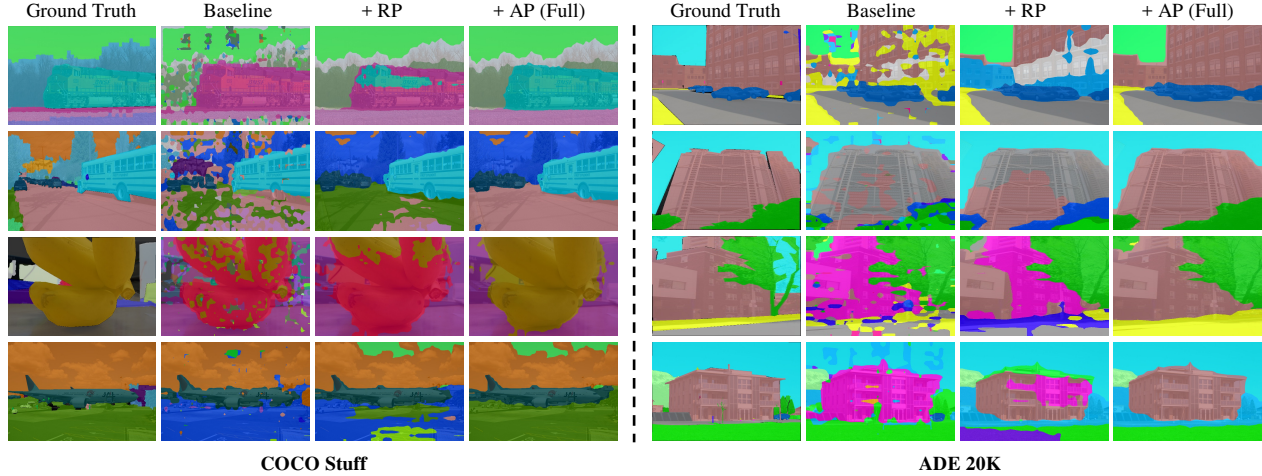


Figure 5. **Visualization of COCO Stuff and ADE20k dataset.** Our FreeCP can effectively eliminate redundancy and resolve ambiguity.

Table 4. **Ablation of textual description on ADE dataset.** We show the results of using different text for initial input and AP.

	Initial	AP	MaskCLIP	SCLIP	GEM	ClearCLIP
Baseline	Template	N/A	13.0	17.2	16.6	17.3
	Vicuna	N/A	13.3	17.2	17.0	17.9
	GPT-3.5	N/A	12.6	16.4	16.6	17.0
FreeCP	Template	Template	17.3	18.4	17.6	18.8
	Template	Vicuna	17.6	18.4	17.8	18.9
	Vicuna	Vicuna	17.5	18.2	18.2	19.4
	Template	GPT-3.5	17.5	18.2	17.6	18.6
	GPT-3.5	GPT-3.5	17.0	17.8	17.0	18.4

Effect of textual description. Tab. 4 evaluates the impact of textual description on segmentation performance. In general, fine-grained descriptions consistently and significantly improve the performance of each method. The descriptions generated by Vicuna perform slightly better than those from GPT [1]. A key observation is that FreeCP outperforms the baseline regardless of the description used. This shows that the improvement is not solely due to the advantages of large language models (LLMs). Additionally, all methods achieve the best results when using Vicuna in the AP stage, while their preferences vary in initial textual selection.

Effect of affinity features. We conduct an ablation study on the choice of affinity features used during refinement. As shown in Tab. 5, using class-agnostic masks from MaskFormer [14] as image self-affinity features yields suboptimal results, as these masks fail to capture contextual relationships between instances. Similarly, although DINO [8] and SAM [27] effectively extract object-level features, they are less capable of modeling intra-class relationships.

Effect on Different Architectures. Our method is compatible with various CLIP frameworks, including both ViT- and

Table 5. **Ablation studies on refinement with different features.**

	Refine	VOC21	PC60	Object	City	ADE	Stuff
MaskFormer		43.5	25.2	23.9	25.8	12.3	17.3
	DINO	61.2	33.0	33.3	30.6	16.4	22.3
	SAM	57.3	31.5	30.8	29.7	15.2	21.2
	CLIP	65.8	35.3	37.2	33.3	18.4	24.9

Table 6. **Ablation studies on different architectures.**

	Methods	VOC21	PC60	Object	City	ADE	Stuff
ViT-L/14	Baseline	46.0	25.8	25.9	27.9	15.0	19.9
	FreeCP	58.9	31.5	33.2	30.1	17.9	22.6
OpenCLIP ViT-L/14	Baseline	27.4	24.9	19.2	26.4	16.0	19.5
	FreeCP	57.5	30.7	32.2	27.7	18.5	22.5
R50x16	Baseline	38.4	18.5	23.7	17.8	11.8	15.2
	FreeCP	52.6	27.7	30.2	20.5	13.0	17.6

ResNet-based architectures. As shown in Tab. 6, FreeCP consistently yields substantial improvements across different backbones. Since ResNet lacks attention layers, we employ DINO features as the affinity representation for the refinement process.

5. Conclusions

In this paper, we identify the core issues as class redundancy and visual-language ambiguity, often arising from the overcomplete vocabulary. Based on these insights, we propose a novel training-free method, FreeCP, designed to purify classes and address these challenges. Extensive experiments across eight benchmarks demonstrate that our method achieves state-of-the-art performance.

Acknowledgments

This project is supported by the National Natural Science Foundation of China (12326618, 62206316), the Project of Guangdong Provincial Key Laboratory of Information Security Technology (2023B1212060026), the Major Key Project of PCL (PCL2024A06), and Alibaba Innovative Research Program. Besides, the authors would like to thank Pengze Zhang (ByteDance), Shuyang Sun (Google DeepMind) and Philip Torr (University of Oxford) for their constructive assistance.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8
- [2] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4981–4990, 2018. 4
- [3] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4253–4262, 2020. 6
- [4] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 3, 6, 7
- [5] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 3, 6, 7
- [6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1209–1218, 2018. 1, 6
- [7] Kaixin Cai, Pengzhen Ren, Yi Zhu, Hang Xu, Jianzhuang Liu, Changlin Li, Guangrun Wang, and Xiaodan Liang. Mixreorg: Cross-modal mixed patch reorganization is a good mask learner for open-world semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 1196–1205, 2023. 3
- [8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 8
- [9] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11165–11174, 2023. 1, 3
- [10] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 1
- [11] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In *Int. Conf. Comput. Vis.*, pages 699–710, 2023. 3
- [12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 1
- [13] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4288–4298, 2022. 1
- [14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1290–1299, 2022. 1, 8
- [15] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 6
- [16] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryoung Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 3
- [17] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3213–3223, 2016. 6
- [18] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11583–11592, 2022. 3
- [19] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. In *Int. Conf. Mach. Learn.*, pages 8090–8102, 2023. 3
- [20] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.*, 111(1):98–136, 2015. 6
- [21] Siqi Fan, Fenghua Zhu, Zunlei Feng, Yisheng Lv, Mingli Song, and Fei-Yue Wang. Conservative-progressive collaborative learning for semi-supervised semantic segmentation. *IEEE Trans. Image Process.*, 2023. 1
- [22] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *Eur. Conf. Comput. Vis.*, pages 540–557. Springer, 2022. 1
- [23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Int. Conf. Mach. Learn.*, pages 4904–4916. PMLR, 2021. 1

- [24] Luo Jiayun, Siddhesh Khandelwal, Leonid Sigal, and Boyang Li. Plug-and-play, dense-label-free extraction of open-vocabulary semantic segmentation from vision-language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 3
- [25] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *Eur. Conf. Comput. Vis.*, 2024. 2, 6, 7
- [26] Laurynas Karazijav, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *Eur. Conf. Comput. Vis.*, 2024. 1, 3, 6, 7
- [27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–4026, 2023. 3, 8
- [28] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *Eur. Conf. Comput. Vis.*, 2024. 1, 3, 6, 7
- [29] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *Eur. Conf. Comput. Vis.*, 2024. 1, 3, 6, 7
- [30] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *Int. Conf. Learn. Represent.*, 2022. 3
- [31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Int. Conf. Mach. Learn.*, 2022. 3
- [32] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 1, 3
- [33] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15305–15314, 2023. 2, 4, 6
- [34] Quande Liu, Youpeng Wen, Jianhua Han, Chunjing Xu, Hang Xu, and Xiaodan Liang. Open-world semantic segmentation via contrasting and clustering vision-language embedding. In *Eur. Conf. Comput. Vis.*, pages 275–292. Springer, 2022. 1
- [35] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *Int. Conf. Mach. Learn.*, pages 23033–23044. PMLR, 2023. 1, 3
- [36] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 891–898, 2014. 6
- [37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafranec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. 3
- [38] 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 *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021. 1, 3, 6, 7
- [39] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *Int. Conf. Comput. Vis.*, pages 5571–5584, 2023. 3
- [40] Pengzhen Ren, Changlin Li, Hang Xu, Yi Zhu, Guangrun Wang, Jianzhuang Liu, Xiaojun Chang, and Xiaodan Liang. Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. In *Int. Conf. Learn. Represent.*, 2023. 3
- [41] Pitchaporn Rewatbowornwong, Nattanat Chatthee, Ekapol Chuangsuwanich, and Supasorn Suwajanakorn. Zero-guidance segmentation using zero segment labels. In *Int. Conf. Comput. Vis.*, pages 1162–1172, 2023. 2, 6, 7
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022. 3
- [43] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 16846–16855, 2022. 1, 2, 4
- [44] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *Eur. Conf. Comput. Vis.*, 2024. 3, 6, 7
- [45] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *Adv. Neural Inform. Process. Syst.*, pages 33754–33767, 2022. 3, 6, 7
- [46] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 3, 6, 7
- [47] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *Eur. Conf. Comput. Vis.*, 2024. 1, 2, 3, 6, 7
- [48] Yuan Wang, Rui Sun, Naisong Luo, Yuwen Pan, and Tianzhu Zhang. Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 6, 7
- [49] Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 3
- [50] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. In *Int. Conf. Learn. Represent.*, 2024. 3

- [51] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *Eur. Conf. Comput. Vis.*, 2024. 3
- [52] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Adv. Neural Inform. Process. Syst.*, pages 12077–12090, 2021. 1
- [53] Yun Xing, Jian Kang, Aoran Xiao, Jiahao Nie, Ling Shao, and Shijian Lu. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. In *Adv. Neural Inform. Process. Syst.*, 2024. 3
- [54] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18134–18144, 2022. 1
- [55] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2935–2944, 2023. 3
- [56] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2955–2966, 2023. 3
- [57] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *Eur. Conf. Comput. Vis.*, pages 736–753. Springer, 2022. 1
- [58] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2945–2954, 2023. 3
- [59] Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. A simple framework for text-supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7071–7080, 2023. 3
- [60] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *Adv. Neural Inform. Process. Syst.*, 2024. 3
- [61] Fei Zhang, Tianfei Zhou, Boyang Li, Hao He, Chaofan Ma, Tianjiao Zhang, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation. In *Adv. Neural Inform. Process. Syst.*, 2024. 3
- [62] Pengze Zhang, Hubery Yin, Chen Li, and Xiaohua Xie. Formulating discrete probability flow through optimal transport. In *Adv. Neural Inform. Process. Syst.*, 2023. 3
- [63] Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, and Luping Zhou. Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23705–23714, 2023. 1
- [64] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2921–2929, 2016. 2, 3
- [65] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *Int. J. Comput. Vis.*, 127:302–321, 2019. 6
- [66] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *Eur. Conf. Comput. Vis.*, pages 696–712, 2022. 1, 3, 6, 7
- [67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 3