

CLIPer: Hierarchically Improving Spatial Representation of CLIP for Open-Vocabulary Semantic Segmentation

Lin Sun¹, Jiale Cao^{1*}, Jin Xie², Xiaoheng Jiang³, Yanwei Pang^{1,4}
¹Tianjin University ²Chongqing University ³Zhengzhou University
⁴Shanghai Artificial Intelligence Laboratory

{sun0806, connor, pyw}@tju.edu.cn, xiejin@cqu.edu.cn, jiangxiaoheng@zzu.edu.cn

Abstract

Contrastive Language-Image Pre-training (CLIP) exhibits strong zero-shot classification ability on image-level tasks, leading to the research to adapt CLIP for open-vocabulary semantic segmentation without training. The key is to improve spatial representation of image-level CLIP, such as replacing self-attention map at last layer with self-self attention map or vision foundation model based attention map. In this paper, we present a novel hierarchical framework, named CLIPer, that hierarchically improves spatial representation of CLIP. The proposed CLIPer includes an early-layer fusion and a fine-grained compensation. We observe that, the embeddings and attention maps at early layers can preserve spatial structural information. Inspired by this, we design the early-layer fusion module to generate segmentation map with better spatial coherence. Afterwards, we employ a fine-grained compensation module to compensate local details using the self-attention maps of diffusion model. We conduct the experiments on eight segmentation datasets. Our CLIPer achieves the state-of-the-art performance on these datasets. With ViT-L and sliding-window inference, CLIPer has the mIoU of 72.2% and 44.7% on VOC and Object, outperforming ProxyCLIP by 11.6% and 5.5%. Our code is available at <https://github.com/linsun449/cliper.code>.

1. Introduction

Open-vocabulary semantic segmentation [2, 50, 52] aims to divide an image into different groups and assign each group a label belonging to arbitrary categories. Recently, the researchers mainly explored to employ vision-language models for open-vocabulary semantic segmentation.

Contrastive Language-Image Pre-training (CLIP) model [26] has shown strong zero-shot capabilities on image-level classification task, due to the pre-training on large-scale

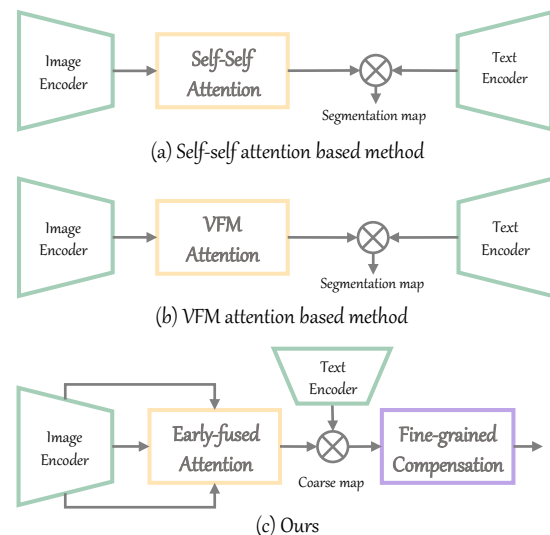


Figure 1. **Comparison with CLIP-based open-vocabulary semantic segmentation approaches without training.** In (a), the approaches [10, 16, 36] replace original self-attention at last layer with self-self attention, which can better maintain spatial coherence. In (b), ProxyCLIP [15] opts for a different strategy, replacing original self-attention map with vision foundation model (VFM) based attention map. In (c), we utilize the embeddings and self-attention maps at early layers to fully exploit spatial information in CLIP. Subsequently, we perform fine-grained compensation using diffusion model as post-processing to improve local details.

image-text paired data [31]. Based on this, several methods have been proposed to adapt CLIP for training-free open-vocabulary semantic segmentation. The key challenge is to improve spatial representation of image-level supervised model for pixel-level segmentation. As shown in Fig. 1(a), some methods modify the original self-attention map at last layer with self-self attention map, better maintaining local spatial information. For instance, MaskCLIP [52] employs an identical self-self matrix as the self-attention map at last layer to generate visual patch embeddings, while SCLIP [36] and ClearCLIP [16] employ the query-to-query or key-

*Corresponding author: Jiale Cao

to-key attention map to replace original self-attention map. Instead of using original or self-self attention map, ProxyCLIP [15] extracts the self-attention map from visual foundation model (VFM) [4] as self-attention map at last layer in Fig. 1(b). These methods enhance segmentation performance of CLIP in open-vocabulary setting without additional training. However, these methods mainly consider improving the self-attention map at last layer of CLIP.

In this paper, we focus on two factors to hierarchically improve spatial representation. (i) The first one is to improve patch-level spatial coherence. We observe that, early layers share similar embedding space with last layer and also have good spatial coherence. This enables training-free feature fusion of early-layer information to improve segmentation. (ii) The second is fine-grained compensation. The patch-level similarity map between image and text is relatively coarse in local details. It is necessary to further improve local details for improved segmentation.

Based on these two factors above, we introduce a novel hierarchical method for open-vocabulary semantic segmentation, named CLIPer. Our CLIPer consists of an early-layer fusion module and a fine-grained compensation module. As shown in Fig. 1(c), the early-layer fusion module integrates patch embeddings and attention maps from early layers to improve spatial coherence of output patch embeddings. Based on the output patch embeddings and text embeddings of arbitrary categories, we can generate the coarse segmentation map. Afterwards, the fine-grained compensation module integrates fine spatial information of Stable Diffusion to compensate the local details. We conduct experiments on various segmentation datasets. The contributions and merits are summarized as

- We propose a novel training-free CLIP-based method that hierarchically improves the spatial representation of CLIP for open-vocabulary semantic segmentation.
- An early-layer fusion strategy is introduced to improve patch-level coherence within CLIP by integrating early-layer information.
- A fine-grained compensation module leverages the fine detail information from diffusion model to refine local details lost in CLIP, leading to more precise segmentation without any training.
- Our method has superior performance on eight datasets. With ViT-L and sliding-window inference, it has averaged score of 48.6%, outperforming ProxyCLIP by 5.7%.

2. Related Work

Vision-language pre-training models. Vision-language pre-training models aim to establish relationships between the images and texts. Among these models, CLIP [26] is one of most successful vision-language models, which is trained on a very large-scale image-text paired dataset. Due to the large-scale pre-training, CLIP exhibits strong

zero-shot classification performance on various image-level tasks. OpenCLIP [7] explores to improve CLIP via conducting a comprehensive experimental analysis of scaling laws. To address the challenge of expensive image-text annotations, ALIGN [13] introduces to use large-scale noisy image-text data for model learning.

Open-vocabulary semantic segmentation. In contrast to traditional semantic segmentation [23, 33, 37], open-vocabulary semantic segmentation [21, 27, 32] segments objects belonging to arbitrary categories. The related methods are divided into training-based [5, 24, 38, 46] and training-free [22, 36] approaches. Training-based methods first train a model on a fixed set of categories from a given training dataset and then apply the learned model to segment objects from arbitrary semantic categories. Some training-based approaches [20, 47] follow two-stage pipeline, where the first stage extracts mask proposals, and the second stage assigns semantic labels to mask proposals. For instance, OVSeg [20] first trains a class-agnostic mask proposals using query-based framework Mask2Former [6], and then fine-tunes CLIP to classify cropped and masked images. Some training-based approaches adopt single-stage pipeline. For instance, SAN [49] introduces a side adapter to adapt CLIP for both classification and segmentation. CAT-Seg [8] constructs pixel-level cost map for segmentation. SED [42] introduces a simple encoder-decoder architecture with category early rejection for efficient segmentation. Cascade-CLIP [18] employs multiple independent text-image decoders to extract features from different layers, introducing additional learnable networks.

Compared to training-based approaches, training-free methods aim to directly adapt vision-language models for open-vocabulary semantic segmentation without any training. Most training-free approaches focus on exploring to improve spatial coherence of image-level supervised CLIP. For instance, MaskCLIP [52] removes the self-attention at last layer, and directly employs value embeddings as output embeddings to perform pixel-level segmentation. Instead of removing self-attention, SCLIP [36] and ClearCLIP [16] employ query-to-query or key-to-key attention map to replace original attention map at last layer. CaR [35] adopts a recurrent framework to progressively enhance segmentation. In addition, some researchers have explored to employ diffusion models for open-vocabulary semantic segmentation. For instance, ODISE [47] employs diffusion model to generate mask proposals, and generates the visual embeddings of masks for classification. OVDiff [14] generates support images of arbitrary categories using diffusion model, and extracts the features of prototypes to segment inference images. DiffSegmenter [39] and iSeg [34] exploit self-attention and cross-attention maps from diffusion models for open-vocabulary segmentation.

The recent DINO-based approaches [15, 40] focus on

improving spatial information of CLIP. For instance, CLIP-DINOiser [40] employs DINO features with good spatial properties to guide CLIP feature learning during training, while ProxyCLIP [15] replaces original self-attention in CLIP with proxy attention in DINO. Our method is related to these approaches, but has significant differences. Instead of integrating spatial coherence of DINO features via training guidance or self-attention replacement, our method employs diffusion model as a post-processing step to refine coarse segmentation and significantly outperforms these approaches with DINO. In addition, our early-layer fusion strategy can fully mine spatial information within CLIP.

3. Methodology

3.1. Preliminary

CLIP. CLIP [26] contains an image encoder and a text encoder. The image encoder has a series of transformer blocks [11], where input image is divided into patches and processed through these blocks. The transformer block contains a residual attention and a residual FFN. Initially, a class token is added to aggregate information from all patches, forming a global representation of image. Subsequently, each transformer block processes input embeddings $F = [F_{cls}, F_1, \dots, F_{hw}]$, where F_{cls} represents the embeddings of class token, and others represent the embeddings of patch tokens. The output embeddings of class token are finally aligned with the embeddings of text encoder.

Stable Diffusion. Image diffusion model generates images starting from random Gaussian noise through a series of denoising steps. By training on large-scale dataset, the diffusion model Stable Diffusion [29] is able to generate high-quality images with rich details. It has been shown that, the features in Stable Diffusion are able to accurately capture local detail information. Therefore, we explore to use Stable Diffusion to improve local details of segmentation.

Attention mechanism. Both CLIP and Stable Diffusion leverage attention mechanism. CLIP employs the self-attention to model relationships between image patches. In contrast, Stable Diffusion incorporates both self-attention and cross-attention, where self-attention extracts spatial coherence within the image, and cross-attention allows the model to incorporate conditioning information (*e.g.*, text description) to guide image generation. The output, whether in self-attention or cross-attention, is calculated by the query Q , key K and value V as follows

$$\text{Att}(Q, K, V) = A \times V, \quad (1)$$

the attention map A is given by

$$A = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right), \quad (2)$$

where d is feature dimensionality. In self-attention, the query, key, and value come from the same embeddings of

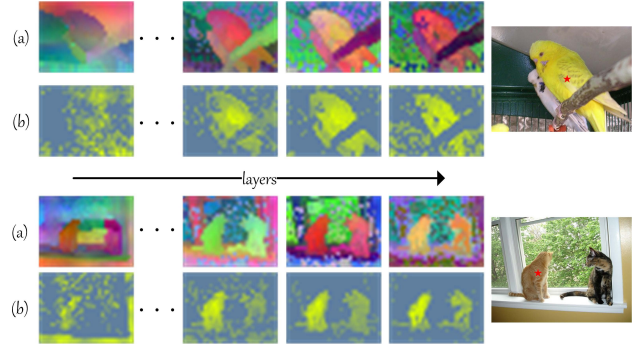


Figure 2. **Visualization of patch embeddings in CLIP.** In (a), we visualize the embeddings using t-SNE, where both the early and later embeddings exhibit good spatial coherence. In (b), we evaluate the cosine similarity between the embeddings of early layers at a specific point and the embeddings at last layer, revealing that the earlier and last embeddings share similar embedding space.

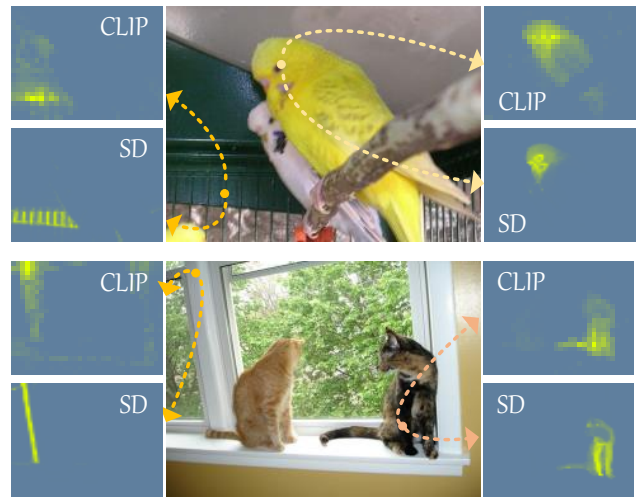


Figure 3. **Visualization of self-attention maps in CLIP and Stable Diffusion (SD).** We show the self-attention maps at selected points for both CLIP and SD. Compared to that of CLIP, the self-attention maps of SD focus more on capturing local details.

image. In contrast, in cross-attention, the query comes from image, while the key and value come from text description.

3.2. Motivation

To adapt pre-trained CLIP for open-vocabulary segmentation, one straightforward approach is to discard the class token and use only the patch tokens to generate pixel-level similarity map with text embeddings. However, since CLIP is pre-trained on image-level classification task, this simple approach usually achieves poor segmentation due to weak spatial coherence of patch embeddings. To address this issue, some approaches [10, 36] primarily focus on modifying the last layer to improve spatial coherence. In contrast, we propose to hierarchically improve spatial representation for

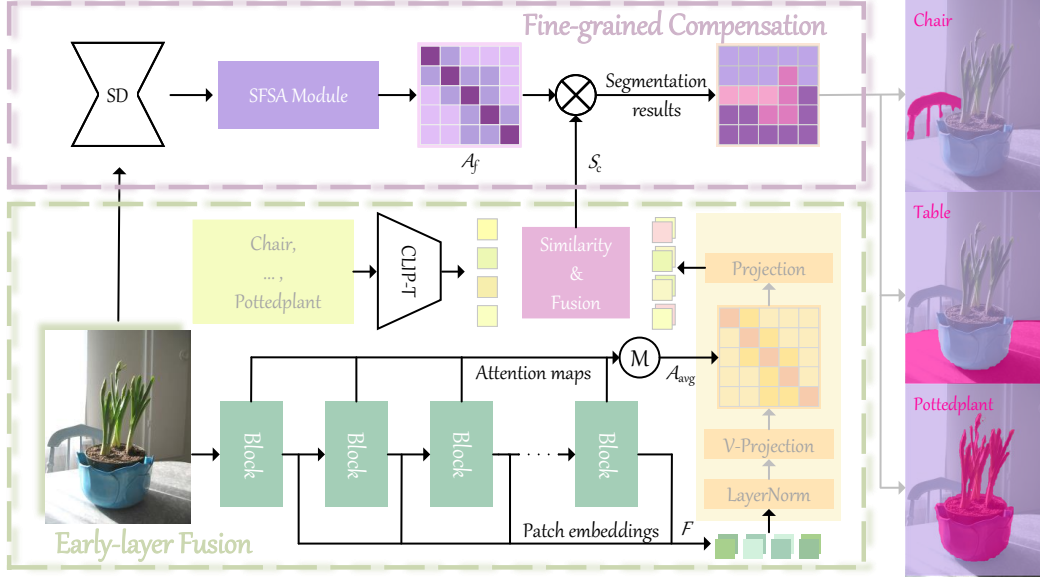


Figure 4. **Overall architecture of our method CLIPer.** Our CLIPer contains two components: early-layer fusion and fine-grained compensation. In the early-layer fusion, we aggregate early-layer information of CLIP image encoder, including embeddings and attention maps, to improve spatial coherence of output embeddings, which are used to generate coarse segmentation map with text embeddings. The fine-grained compensation aims to employ self-attention maps of Stable Diffusion to refine local details of coarse segmentation map.

better segmentation inspired by two aspects of observations.

The first is, the embeddings at early layers can be used to improve spatial coherence. First, as shown in Fig. 2(a), similar to that at last layer, the patch embeddings at early layers also retain spatial information. Second, in Fig. 2(b), early-layer embeddings share similarities with the embeddings at last layer. This similarity enables effective fusion of early- and late-layer information without training.

The second is, the pre-trained Stable Diffusion (SD) can generate high-quality images with rich details. As in Fig. 3, by visualizing the self-attention maps in CLIP and SD, we observe that the attention maps of SD effectively capture local details. This contrasts with the attention maps in CLIP, which typically respond to broader semantic areas. This key observation suggests that we can integrate the fine spatial information from SD to improve coarse segmentation generated by CLIP.

3.3. Framework

Overview. Inspired by motivation and insights above, we propose a novel hierarchical approach for open-vocabulary semantic segmentation. Fig. 4 presents an overall architecture of our proposed method, named CLIPer. Our CLIPer consists of two complementary components. In first component, we leverage an early-layer fusion module to generate patch embeddings with better spatial coherence, and then generate coarse segmentation map according to the similarity between patch embeddings and text embeddings of text encoder. In second component, we perform fine-grained

compensation using the attention maps of Stable Diffusion.

Early-layer fusion. This module aims to improve spatial coherence using the embeddings of early layers. Specifically, given an image, we first divide the image into patch embeddings $F^0 \in \mathbb{R}^{(hw+1) \times D}$, and then feed these embeddings to a series of transformer blocks. For n -th transformer block, we generate the query Q^n , key K^n , and value V^n as $Q^n = \text{Proj}_q(\text{LN}(F^{n-1}))$, $K^n = \text{Proj}_k(\text{LN}(F^{n-1}))$, $V^n = \text{Proj}_v(\text{LN}(F^{n-1}))$. Here, LN denotes layer normalization, and the projections Proj_q , Proj_k , and Proj_v respectively present a linear layer. With the query Q^n , key K^n , and value V^n , the output embeddings F^n of n -th transformer block are calculated by

$$\bar{F}^n = \text{Att}(Q^n, K^n, V^n) + F^{n-1}, \quad (3)$$

$$F^n = \text{FFN}(\text{LN}(\bar{F}^n)) + \bar{F}^n, \quad (4)$$

where FFN stands for a feed-forward network. The attention map of n -th transformer, which is generated in attention operation of Eq. 3, denoted as A^n .

Similarly, we can generate the embeddings and attention maps of all transformer blocks up to the penultimate layer, denoted as two sets: $\mathcal{F} = \{F^i | i = 1, 2, \dots, N-1\}$ and $\mathcal{A} = \{A^i | i = 1, 2, \dots, N-1\}$. We first generate an averaged attention map as

$$A_{avg} = \frac{1}{N} \sum_{n=1}^{N-1} A^n. \quad (5)$$

Then, we replace the original self-attention map at last layer with the averaged attention map A_{avg} , and then feed all the embeddings to the last layer. Similar to ClearCLIP [16], we omit the feed-forward network and residual connections in the last transformer block, which can simplify the representation while aligning text embeddings better. As a result, we generate multiple output embeddings for different layers.

Lastly, we compute the cosine similarity between multiple output embeddings and text embeddings derived from the CLIP text encoder, and calculate the averaged similarity map. This averaged similarity map is used as the coarse segmentation by mapping each patch embedding to the candidate category embeddings.

Fine-grained compensation. The patch-level segmentation generated by CLIP remains relatively coarse, limiting segmentation accuracy. To better compensate local details of coarse map, we employ self-fused self-attention (SFSA) to leverage self-attention maps from Stable Diffusion, as we find it is effective at capturing fine-grained local information. This locality-preserving characteristic is highly beneficial for refining spatial details of patch-level segmentation.

Specifically, we first feed the image along with an empty (null) textual prompt into Stable Diffusion, obtaining the corresponding multi-head self-attention maps at the highest spatial resolutions. We denote the attention maps as $A_m \in \mathbb{R}^{H \times L \times L}$, where H represents the number of attention heads, and L indicates the spatial size of feature maps in Stable Diffusion. Then, we fuse these attention maps A_m by matrix chain multiplication across the attention heads, which is formulated as

$$A_f = A_m[0] \times A_m[1] \times \cdots \times A_m[H - 1], \quad (6)$$

where $A_m[i]$ means the i -th head self-attention map. Afterwards, we utilize the fused attention map A_f to refine the upscaled coarse segmentation map S_c as

$$S_f = A_f \times S_c. \quad (7)$$

Finally, we upscale S_f to the resolution of input image, yielding the fine-grained pixel-level segmentation map.

4. Experiment

4.1. Experimental Setups

Datasets. We evaluate CLIPer on eight datasets: (1) Considering the background category. We use PASCAL VOC (VOC) [12], PASCAL Context (Context) [25], and COCO Object (Object) [3]; (2) Without considering the background category. We use PASCAL VOC (VOC20), PASCAL Context (Context59), COCO-Stuff (Stuff) [3], Cityscapes [9], and ADE20K (ADE) [51]. For performance evaluation, we use the validation set from each dataset. In addition, for weakly supervised semantic segmentation, we

evaluate the pseudo-mask generation performance on the training sets of VOC and COCO datasets.

Metrics. We use mean Intersection over Union (mIoU) to evaluate segmentation, and adopt mAP, F1 score, Precision (P), and Recall (R) to evaluate image-level classification.

Implementation details. We implement our method on a single RTX 3090 with 24G memory. We employ ViT-B and ViT-L as the backbones, and use Stable Diffusion V2.1 [29] for fine-grained compensation. In the Stable Diffusion, we extract the attention maps at time-step 45 in total of 50 steps. We set text prompts including category descriptions similar to SCLIP [36] and ProxyCLIP [15]. We resize all the input images to a shorter side of 336 pixels while maintaining the original aspect ratio, similar to ProxyCLIP [15]. We directly feed the entire image into the CLIP image encoder for efficiency, and also report the results using a sliding window strategy similar to [52], [36], [16] and [15].

4.2. Comparison With Other Methods

On mIoU. Table 1 compares our method with some state-of-the-art methods on various datasets. Our method almost achieves the best averaged performance on all these datasets when using both ViT-B and ViT-L backbones. For instance, on VOC with the ViT-L backbone, SCLIP [36] has the mIoU score of 43.5%, ProxyCLIP [15] has the mIoU score of 60.6%, while our method achieves the mIoU score of 69.8% without sliding-window strategy. Namely, our method outperforms SCLIP and ProxyCLIP by 26.3% and 9.2% on VOC. On ADE with the ViT-L backbone, ClearCLIP [16] and ProxyCLIP achieve the mIoU scores of 15.0% and 22.6%, while our method achieves the mIoU score of 24.4%. Namely, our method has the improvements of 9.4% and 1.8% on ADE. With sliding-window inference, it has better performance.

On category classification and mask prediction. Open-vocabulary semantic segmentation can be viewed as two aspects: category classification and mask prediction. To deeply show the advantage of our proposed method on these two aspects, we provide more comparisons with other methods via two experiments. (i) We present image-level precision (P) and recall (R) comparison to show the advantages of identifying the categories within image. (ii) We compare the segmentation accuracy when giving image-level category labels, demonstrating the advantages of mask prediction. Weakly supervised semantic segmentation aims to train the model based on image-level category labels of training set. By comparing our method with weakly supervised approaches, it can demonstrate the benefits of our method in mask prediction.

In Table 2, we calculate image-level classification scores for all methods, and report mAP, F1, P, and R. Our method achieves best performance on all these metrics. We also observe similar improvement with ViT-B. It demonstrates

Type	Method	Encoder	VOC	Context	Object	VOC20	Context59	Stuff	Cityscapes	ADE	Avg.
Training-based (weakly-supervised)	SegCLIP [24]	ViT-B/16	52.6	24.7	26.5	-	-	-	-	-	-
	ViewCo [28]	ViT-S/16	52.4	23.0	23.5	-	-	-	-	-	-
	OVSegmentor [46]	ViT-B/16	53.8	20.4	25.1	-	-	-	-	-	-
	CoCu [44]	ViT-S/16	51.4	23.6	22.7	-	-	15.2	22.1	12.3	-
	GroupViT [45]	ViT-S/16	50.4	18.7	27.5	79.7	23.4	15.3	11.1	9.2	29.4
	TCL [5]	ViT-B/16	51.2	24.3	30.4	77.5	30.3	19.6	23.1	14.9	33.9
	CLIP-DINOiser [40]	ViT-B/16	62.2	32.4	35.0	80.9	35.9	24.6	31.7	20.0	40.3
Training-free	CLIP [‡] [26]	ViT-B/16	16.4	8.4	5.6	41.9	9.2	4.4	5.5	2.9	11.8
	CLIPsurgery [19]	ViT-B/16	-	29.3	-	-	-	21.9	31.4	-	-
	MaskCLIP [‡] [52]	ViT-B/16	38.8	23.6	20.6	74.9	26.4	16.4	12.6	9.8	27.9
	SCLIP [‡] [36]	ViT-B/16	59.1	30.4	30.5	80.4	34.2	22.4	32.2	16.1	38.2
	ClearCLIP [‡] [16]	ViT-B/16	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	38.1
	ProxyCLIP [‡] [15]	ViT-B/16	61.3	35.3	37.5	80.3	39.1	26.5	38.1	20.2	42.3
	CLIPer (Ours)	ViT-B/16	65.9	37.6	39.0	85.2	41.7	27.5	38.3	21.4	44.4
	CLIPer[‡] (Ours)	ViT-B/16	66.5	38.3	40.0	86.0	42.4	28.6	38.7	22.0	45.3
	CLIP [‡] [26]	ViT-L/14	8.2	4.1	2.7	15.6	4.4	2.4	2.5	1.7	5.2
	MaskCLIP [‡] [52]	ViT-L/14	23.3	11.7	7.2	29.4	12.4	8.8	11.5	7.2	13.9
	SCLIP [‡] [36]	ViT-L/14	43.5	22.3	25.0	69.1	25.2	17.6	18.6	10.9	29.0
	ClearCLIP [‡] [16]	ViT-L/14	46.1	26.7	30.1	80.0	29.6	19.9	27.9	15.0	34.4
	CaR [35]	ViT-L/14	67.6	30.5	36.6	91.4	39.5	-	-	17.7	-
	ProxyCLIP [‡] [15]	ViT-L/14	60.6	34.5	39.2	83.2	37.7	25.6	40.1	22.6	42.9
	ProxyCLIP [‡] [15]	ViT-H/14	65.0	35.4	38.6	83.3	39.6	26.8	42.0	24.2	44.4
CLIPer (Ours)	ViT-L/14	69.8	38.0	43.3	90.0	43.6	28.7	41.6	24.4	47.3	
CLIPer[‡] (Ours)	ViT-L/14	72.2	39.5	44.7	89.8	44.6	30.4	42.5	25.0	48.6	

Table 1. **Comparison with existing open-vocabulary segmentation methods.** For training-free approaches, [‡] denotes using sliding-window inference. Among training-free approaches, our method, even without sliding-window inference, has best averaged performance.

Method	Encoder	VOC				Context				Object			
		mAP	F1	P	R	mAP	F1	P	R	mAP	F1	P	R
MaskCLIP [52]	ViT-L/14	87.3	63.2	56.3	72.1	58.6	48.9	48.5	49.3	67.4	48.2	47.6	52.4
SCLIP [36]	ViT-L/14	92.6	77.3	85.5	70.5	61.1	54.8	61.9	49.1	74.4	56.6	66.6	49.2
ClearCLIP [16]	ViT-L/14	91.7	76.1	85.7	68.5	60.9	54.9	52.8	57.2	73.4	53.5	61.3	47.5
ProxyCLIP [15]	ViT-L/14	94.0	77.6	84.5	71.7	64.1	56.1	53.7	58.8	77.7	60.5	68.2	54.3
CLIPer (Ours)	ViT-L/14	94.6	86.0	86.7	85.3	68.9	63.3	63.4	64.3	77.9	62.3	69.7	56.3

Table 2. **Comparison of image-level category classification capability with existing methods.** We calculate classification scores of different categories by max-pooling segmentation maps, and then calculates the results of mAP, F1, Precision (P), and Recall (R). Our method achieves best results across all datasets, demonstrating its superior performance on category classification.

that, our method performs better on category classification, which is useful for open-vocabulary semantic segmentation.

Table 3 further compares our method with weakly supervised semantic segmentation approaches for pseudo mask generation, where image-level category labels are given. Compared to these approaches, our method achieves best performance. For instance, our method outperforms CLIP-ES [21] and iSeg [34] by 6.1% and 1.7%. It demonstrates that, our method can improve mask prediction, which is important for open-vocabulary semantic segmentation.

Inference time. Table 4 compares inference time and accuracy. We present various versions to accommodate different needs. Compared to ClearCLIP [16], our CLIPer* (w/o fine-grained compensation) has faster speed and higher mIoU. Compared to ProxyCLIP [15], our CLIPer* has faster speed and comparable mIoU. Our CLIPer with fine-grained compensation further improves CLIPer*.

Qualitative results. Fig. 5 presents some examples of qualitative comparison on VOC, Context, and Object. Our proposed method has more accurate segmentation maps and precise classification, compared to these methods [10, 15, 16]. For instance, our method has finer segmentation on bicycle and correct classification of person in second column, and accurate segmentation on sofa in fifth column.

4.3. Ablation Study

Impact of different modules. Table 5 presents the results of integrating different modules. The baseline replaces the original self-attention map at last layer with value-to-value attention map, and removes the feed-forward network (FFN) and residual connection. The baseline achieves the mIoU scores of 51.2%, 26.5%, and 32.3% on VOC, Context, and Object. When adding early-layer fusion (ELF) module, it has the mIoU scores of 61.2%, 34.3%, and 39.6%

Type	Method	VOC	COCO
Training-based	IRN [1]	66.5	42.4
	AdvCAM [17]	55.6	35.8
	MCTformer [48]	61.7	-
	ToCo [30]	72.2	-
	CLIMS [43]	56.6	-
Training-free	CLIP-ES [21]	70.8	39.7
	DiffSegmenter [39]	70.5	-
	T2M [41]	72.7	43.7
	iSeg [34]	75.2	45.5
	CLIPer (Ours)	76.9	47.3

Table 3. **Comparison with weakly supervised semantic segmentation approaches.** Both our method and these weakly supervised approaches predict pseudo masks according to given image-level category labels. Our CLIPer achieves the best results, showing that our method has better results on mask prediction.

Method	Encoder	Input size	Time(ms) ↓	mIoU ↑
ClearCLIP [‡] [16]	ViT-B/16	~ 448 × 624	22	51.8
ProxyCLIP [‡] [15]	ViT-B/16	~ 336 × 468	82	61.3
CLIPer* (Ours)	ViT-B/16	~ 336 × 468	14	60.1
CLIPer (Ours)	ViT-B/16	~ 336 × 468	158	65.9
CLIPer[‡] (Ours)	ViT-B/16	~ 336 × 468	186	66.5
ClearCLIP [‡] [16]	ViT-L/14	~ 448 × 624	68	46.1
ProxyCLIP [‡] [15]	ViT-L/14	~ 336 × 468	105	60.6
CLIPer* (Ours)	ViT-L/14	~ 336 × 468	47	61.2
CLIPer (Ours)	ViT-L/14	~ 336 × 468	192	69.8
CLIPer[‡] (Ours)	ViT-L/14	~ 336 × 468	257	72.2

Table 4. **Comparison in terms of mIoU and inference time on VOC.** * denotes not using fine-grained compensation, and ‡ denotes sliding-window inference. Our CLIPer* has the fastest speed, while our CLIPer and CLIPer[‡] has the best performance.

ELF	FGC	VOC	Context	Object
✗	✗	51.2	26.5	32.3
✓	✗	61.2	34.3	39.6
✗	✓	62.8	29.7	36.4
✓	✓	69.8	38.0	43.3

Table 5. **Ablation study of different modules in our CLIPer.** ELF represents early-layer fusion, and FGC represents fine-grained compensation. Our proposed modules can significantly improve the performance of the baseline.

on VOC, Context, and Object, outperforming the baseline by 10.0%, 7.8%, 7.3%. When only using fine-grained compensation (FGC) module, it outperforms the baseline by 11.6%, 3.2%, 4.1%. When integrating the EFL and FGC modules together, it totally has the improvements of 18.6%, 11.5%, and 11.0% on three datasets, respectively. This significantly demonstrates that, our proposed modules can improve open-vocabulary segmentation performance.

Effect of early-layer fusion. Table 6 compares our early-layer fusion with some self-self attention operations. Our early-layer fusion module fuses both the patch embeddings of early layers and the attention maps of early layers. In the

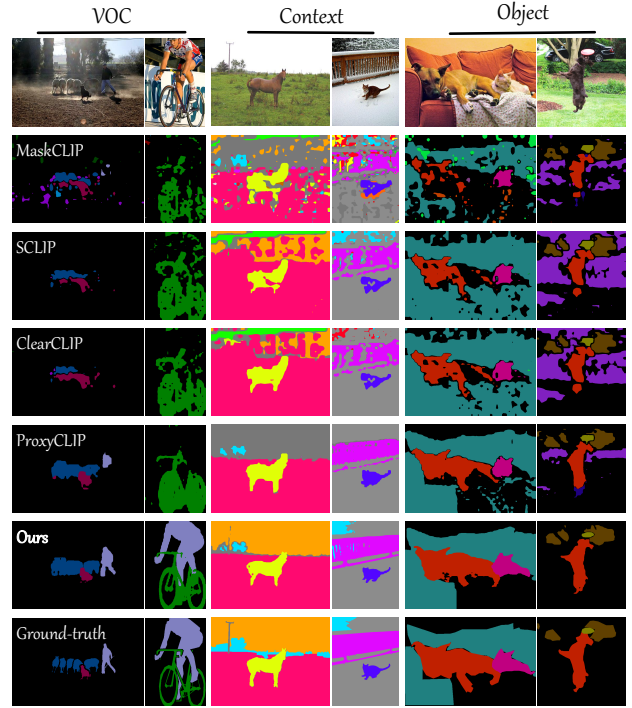


Figure 5. **Qualitative comparison with existing methods.** We show the segmentation results on three different datasets. Compared to these methods, our method has more accurate segmentation results which are closer to the ground-truths.

Attention Type	Early-layer Embeddings	VOC	Context	Object
Query-query	✗	50.1	22.5	28.7
Key-key	✗	44.8	25.4	27.0
Value-value	✗	51.2	26.5	32.3
Identity matrix	✗	40.1	22.0	25.7
Early-layer attention	✗	58.6	33.1	38.4
Query-query	✓	55.7	26.9	32.0
Key-key	✓	53.2	29.5	33.1
Value-value	✓	54.8	29.4	34.4
Identity matrix	✓	43.3	24.8	27.8
Early-layer attention	✓	61.2	34.3	39.6

Table 6. **Impact of different designs in our early-layer fusion.** Our early-layer fusion contains fusing the embeddings and attention maps. In the top part, we compare our fused attention with some self-self attention operations. In the bottom part, we feed the early-layer embeddings to the attention map at last layer.

top part, we compare our early-layer fused attention with these self-self attention operations, which all take the embeddings at last layer as input. Compared to these self-self attention operations, our early-layer fused attention has the best performance. For instance, on VOC, Context, and Object, the query-query attention achieves the mIoU scores of 50.1%, 22.5%, and 28.7%, while our early-layer fused attention has the mIoU scores of 58.6%, 33.1%, and 38.4%. Namely, our early-layer fused attention outperforms query-query attention by 8.5%, 10.6%, and 9.7%, respectively.

Type	VOC	Context	Object
Single	65.3	36.1	41.2
Mean	64.9	36.1	41.0
Multiplication	69.8	38.0	43.3

Table 7. **Comparison of different strategies using the multi-head attention maps in Stable Diffusion.** Single represents that, we evaluate each single head and report the best-performing head. Mean represents that, we average multi-head attention maps, and multiplication presents that, we perform matrix multiplication to fuse multi-head attention maps.

Strategy	VOC	Context	Object
Our ELF (<i>i.e.</i> , CLIPer*)	61.2	34.3	39.6
With attention from DINO	68.2	35.8	41.9
With attention from diffusion model	69.8	38.0	43.3

Table 8. **Comparison with proxy attention.** We replace diffusion model with proxy attention in our FGC.

In bottom part, we show the impact of feeding early-layer embeddings to last layer. When integrating early-layer embeddings into early-layer attention, it has the improvements of 2.6%, 1.2%, and 1.2% on VOC, Context, and Object, respectively. We observe that, using early-layer embeddings can also improve the performance of existing self-attention operations. For instance, when combining it with value-value attention, it has the improvements of 3.6%, 2.9%, and 2.1% on VOC, Context, and Object, respectively.

Effect of fine-grained compensation. Table 7 presents different strategies to fuse multi-head attention maps in Stable Diffusion (SD), including a single head (Single), averaging all heads (Mean), and combining all heads with matrix multiplication (multiplication). Compared to the baseline, all three strategies can improve the performance, demonstrating that the attention maps in SD can improve CLIP-based segmentation. Among these strategies, matrix multiplication has best performance, which is adopted as final setting.

Table 8 shows the effect of compensating with different models. The more detailed comparison can be found in supplementary materials. Both DINO and diffusion model can improve segmentation, while diffusion model achieves better performance. The reason is that, diffusion model mainly focuses on image generation, naturally performing better in dealing with local details via large-scale pre-training.

Qualitative comparison. Fig. 6 shows some examples before and after fine-grained compensation. Before fine-grained compensation, the attention maps in (b) provide coarse spatial structure information of objects. By using our fine-grained compensation, the attention maps in (c) provide more accurate responses around object contour. As a result, compared to that in (d), using our fine-grained compensation has more accurate segmentation maps in (e). It demonstrates that, our fine-grained compensation can improve local details of coarse segmentation maps.

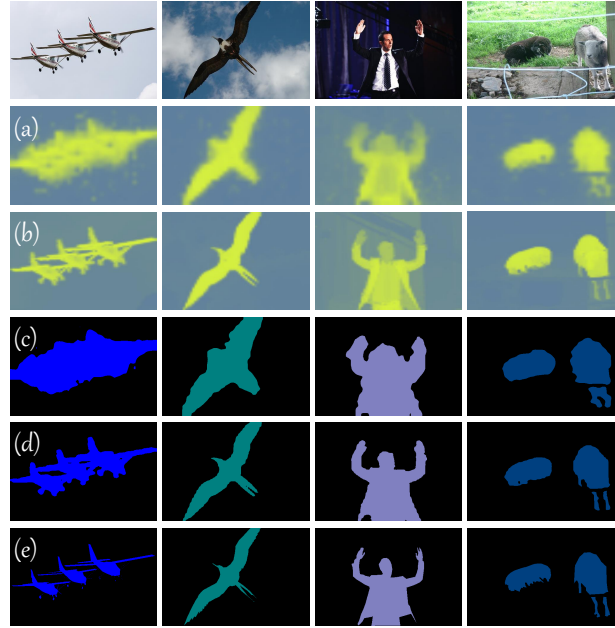


Figure 6. **Visualization of fine-grained compensation.** Given the input images in the top, we present the attention maps before and after fine-grained compensation in (a) and (b), show the binary segmentation maps in (c) and (d). We also present the ground-truth segmentation maps in (e).

5. Conclusion

This paper presents CLIPer, a novel training-free method to hierarchically improve the spatial representation of CLIP for open-vocabulary semantic segmentation. To achieve this goal, we design two components, including early-layer fusion and fine-grained compensation. The early-layer fusion aims to improve spatial coherence of output patch embeddings by using the early-layer information of patch embeddings and attention maps. The fine-grained compensation module employs the fine attention maps of diffusion model to further improve local details of segmentation maps. Our method achieves superior performance on various public datasets. However, we still observe that, our method struggles from accurately segmenting the tiny objects. In future, we will explore how to adapt the pre-trained model with high-resolution input for improving tiny object segmentation. In addition, the frozen Stable Diffusion pretrained on image-level image-text pairs could be used for improving segmentation, suggesting that it is promising to explore better image-text pretraining strategy especially for high-quality open-vocabulary segmentation.

Acknowledgment: This work was supported by the National Key Research and Development Program of China (No. 2022ZD0160400), Natural Science Foundation of China (No. 62271346, 62206031, U21B2037).

References

- [1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2019. 7
- [2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 468–479, 2019. 1
- [3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocomp: Thing and stuff classes in context. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1209–1218, 2018. 5
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 9650–19660, 2021. 2
- [5] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 2, 6
- [6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2
- [7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 2
- [8] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2023. 2
- [9] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016. 5
- [10] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. In *International Conference on Machine Learning*, pages 8090–8102, 2023. 1, 3, 6
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3
- [12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 5
- [13] 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 *International Conference on Machine Learning*, pages 4904–4916, 2021. 2
- [14] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 299–317, 2024. 2
- [15] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88, 2024. 1, 2, 3, 5, 6, 7
- [16] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ClearCLIP: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160, 2024. 1, 2, 5, 6, 7
- [17] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021. 7
- [18] Yunheng Li, Zhong-Yu Li, Quansheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *International Conference on Machine Learning*, pages 28243–28258, 2024. 2
- [19] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025. 6
- [20] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yanan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2
- [21] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 2, 6, 7
- [22] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. TagCLIP: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *AAAI Conference on Artificial Intelligence*, pages 3513–3521, 2024. 2
- [23] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In

- IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 431–440, 2015. 2
- [24] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044, 2023. 2, 6
- [25] 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 Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 5
- [26] 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 *International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 2, 3, 6
- [27] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *IEEE/CVF International Conference on Computer Vision*, pages 5548–5561, 2023. 2
- [28] 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 *International Conference on Learning Representations*, 2023. 6
- [29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022. 3, 5
- [30] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023. 7
- [31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Ross Wightman Cade Gordon, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, pages 25278–25294, 2022. 1
- [32] Gyungin Shin, Weidi Xie, and Samuel Albanie. ReCo: Retrieve and co-segment for zero-shot transfer. In *Advances in Neural Information Processing Systems*, pages 33754–33767, 2022. 2
- [33] Guolei Sun, Yun Liu, Henghui Ding, Min Wu, and Luc Van Gool. Learning local and global temporal contexts for video semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6919–6934, 2024. 2
- [34] Lin Sun, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. iSeg: An iterative refinement-based framework for training-free segmentation. *arXiv preprint arXiv:2409.03209*, 2024. 2, 6, 7
- [35] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. CLIP as RNN: Segment countless visual concepts without training endeavor. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13171–13182, 2024. 2, 6
- [36] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332, 2024. 1, 2, 3, 5, 6
- [37] Hefeng Wang, Jiale Cao, Jin Xie, Aiping Yang, and Yanwei Pang. Implicit and explicit language guidance for diffusion-based visual perception. *IEEE Transactions on Multimedia*, 27:466–476, 2024. 2
- [38] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 2
- [39] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *arXiv preprint arXiv:2309.02773*, 2023. 2, 7
- [40] Monika Wysockańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. *European Conference on Computer Vision*, pages 320–337, 2024. 2, 3, 6
- [41] Changming Xiao, Qi Yang, Zhou Feng, and Changshui Zhang. From text to mask: Localizing entities using the attention of text-to-image diffusion models. *arXiv preprint arXiv:2309.04109*, 2023. 7
- [42] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3426–3436, 2024. 2
- [43] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4492, 2022. 7
- [44] Yun Xing, Jian Kang, Aoran Xiao, Jiahao Nie, Shao Ling, and Shijian Lu. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, pages 68798–68809, 2023. 6
- [45] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 6
- [46] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2935–2944, 2023. 2, 6
- [47] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2
- [48] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4310–4319, 2022. 7
- [49] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 2
- [50] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *IEEE/CVF International Conference on Computer Vision*, pages 2002–2010, 2017. 1
- [51] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 5
- [52] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712, 2022. 1, 2, 5, 6