ClearCLIP: Decomposing CLIP Representations for Dense Vision-Language Inference

Mengcheng Lan[1], Chaofeng Chen[1], Yiping Ke[2], Xinjiang Wang[3], Litong Feng[3⋆], and Wayne Zhang[3]

[1] S-Lab, Nanyang Technological University
[2] CCDS, Nanyang Technological University [3] SenseTime Research
lanm0002@e.ntu.edu.sg {chaofeng.chen, ypke}@ntu.edu.sg
{wangxinjiang, fenglitong, wayne.zhang}@sensetime.com
https://github.com/mc-lan/ClearCLIP

Abstract. Despite the success of large-scale pretrained Vision-Language Models (VLMs) especially CLIP in various open-vocabulary tasks, their application to semantic segmentation remains challenging, producing noisy segmentation maps with mis-segmented regions. In this paper, we carefully re-investigate the architecture of CLIP, and identify residual connections as the primary source of noise that degrades segmentation quality. With a comparative analysis of statistical properties in the residual connection and the attention output across different pretrained models, we discover that CLIP's image-text contrastive training paradigm emphasizes global features at the expense of local discriminability, leading to noisy segmentation results. In response, we propose ClearCLIP, a novel approach that decomposes CLIP's representations to enhance open-vocabulary semantic segmentation. We introduce three simple modifications to the final layer: removing the residual connection, implementing the self-self attention, and discarding the feed-forward network. ClearCLIP consistently generates clearer and more accurate segmentation maps and outperforms existing approaches across multiple benchmarks, affirming the significance of our discoveries.

Keywords: Semantic segmentation · Vision language model · Open vocabulary

1 Introduction

Large-scale Vision-Language pre-trained Models (VLMs), represented by the Contrastive Language-Image Pre-training (CLIP) family [8,35,43], have demonstrated remarkable generality and robustness across a diverse range of downstream tasks, *e.g.*, zero-shot image classification [19,35], visual question answering [2, 21, 51] and image-text retrieval [9, 24, 33]. There is a growing interest in leveraging the power of CLIP for open-vocabulary and zero-shot problems. However, CLIP falls behind to maintain its zero-shot capabilities for dense prediction

⋆ Corresponding author.

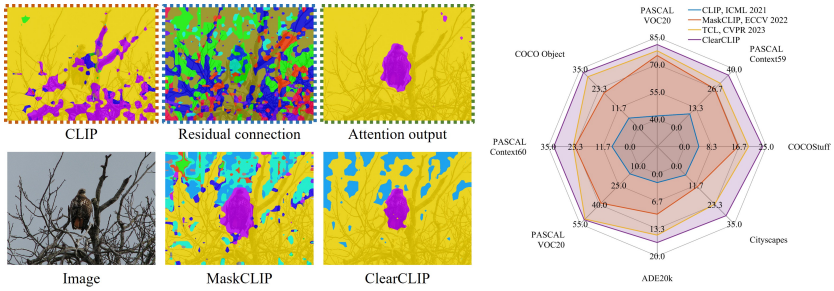


Fig. 1: Left: Example of open-vocabulary semantic segmentation. CLIP [35] fails to localize the object. MaskCLIP [56] can localize the foreground and background but still exhibits significant noise. Our proposed method, ClearCLIP, achieves high-quality segmentation map. Our key insight is that vanilla CLIP's segmentation map can be decomposed into a cluttered map of residual connection and a clearer and smoother map of attention output from the last transformer layer. Right: Comparison of open-vocabulary semantic segmentation performance.

tasks especially semantic segmentation, as depicted by previous works [26, 56]. This limitation arises primarily from the training data used for VLMs, which consists largely of image-level labels and lacks sensitivity to visual localization. For example, as shown in Fig. 1 (top-left image), the segmentation map generated by CLIP based on patch-level cosine similarity between visual and textual features reveals many misclassified patches and significant noise, showcasing the limitation of CLIP in dense visual localization. Recent works [3, 26, 40, 56] usually attribute noisy representations in CLIP to the self-attention layers, and have achieved great progress by improving this module. MaskCLIP [56] adopted a simple technique of setting *query-key* attention map of the last block as an identical matrix, resulting in an improvement in mIoU on COCOStuff [4] from 4.4 to 16.4. CLIPSurgery [26] argued that the *value-value* attention map is cleaner and further improved the mIoU to 21.9. The recent study SCLIP [40] combined the *query-query* and *value-value* attention and achieved even better results. While these enhancements in attention mechanisms lead to prioritization of relevant context and improved performance, their segmentation results still exhibit some noise. Such noise becomes more pronounced and leads to deteriorated performance when larger backbones are applied. These raise the fundamental question: *where do these noises originate and how do they surface in the CLIP models*?

To answer the *where* question, we conduct a thorough investigation into the architecture of CLIP. We are surprised to find that the residual connection, proposed by ResNet and commonly employed in transformer architectures, has a significant effect on the adaptation of CLIP to open-vocabulary semantic segmentation. To elucidate this, we decompose the output of CLIP's vision encoder into two components: the residual connection and the attention output, which is achieved by directly separating them in the last layer of the vision encoder. As illustrated in Fig. 1 (top three images), the segmentation result obtained from

the residual connection exhibits noticeable noise, while the attention output features produce significantly clearer results with superior localization properties. From these observations, we propose that *the noises present in the segmentation map mainly originate from the residual connection.*

To delve into *how* these noises emerge, we begin by comparing the statistical properties of CLIP's residual connection and attention output. Notably, we observe a significant discrepancy in their normalized entropy in CLIP: while the entropy of the residual connection tends to be near 0 along layers, the entropy of the attention output remains at 1. This finding aligns with our observation that its residual connection contains much larger maximum values along layers. Accordingly, the final output of CLIP, *i.e.*, the addition of residual connection and attention output, exhibits similar properties to the residual connection. Our findings also resonate with those of [12], which identify many artifacts with high-norm in the final feature maps of large-scale pretrained models. When we look closely at the residual connection maps, we see that these peak values are concentrated in a few channels. In other words, most feature vectors in the residual connection maps share the same peak dimensions, *i.e.*, similar directions in the latent space. This property makes it difficult to distinguish each spatial feature vector through cosine similarity, thereby contributing to the generation of noises. Conversely, the self-attention mechanism in the attention branch is learned to separate dissimilar spatial features, alleviating such issues. Next, we examine the DINO [5], which is also a transformer architecture but pretrained in a self-supervised way. We find that these two feature maps of DINO do not show such discrepancies in entropy. Therefore, we propose that *the high-level supervision in CLIP emphasizes the global feature direction in the residual latent space, making local feature vectors less distinguishable and leading to noise in residual features.*

Based on these discoveries, we revisit recent methods [26, 40, 56] and find that the performance enhancements observed in these methods can be partially attributed to the reduced influence of the residual connection when the attention output is strengthened. We then deduce that two critical factors play a pivotal role in adapting CLIP for dense vision-language inference: the reduction in the impact of the residual connection and the reorganization of spatial information through the self-self attention. Guided by these insights, we introduce our approach, ClearCLIP, which incorporates three straightforward modifications to the final layer of CLIP: eliminating the residual connection, adopting the self-self attention, and discarding the Feed-Forward Network (FFN). These modifications are designed to boost the attention output, thereby producing a clearer representation for the task of open-vocabulary semantic segmentation, as shown in Fig. 1. Extensive experiments on 8 benchmark datasets demonstrate the effectiveness of ClearCLIP.

2 Related Work

Vision-language pre-training. VLMs has experienced significant progress in recent years. One notable family of vision-language models is based on con-

trastive learning [1,8,19,25,32,35,43,50,53]. Among them, CLIP [35] trained on a private WIT-400M with image-text pairs achieves promising zero-shot capabilities for downstream tasks such as image-text retrieval, image classification via text prompts. ALIGN [19] adopts the same dual-encoder architecture as CLIP but is trained on a private dataset with over one billion noisy image-text pairs. Additionally, OpenCLIP [8] explores scaling laws for CLIP by training the models on the public LAION [37] dataset with up to two billion image-text pairs. Another line of research [7, 22, 24] focuses on shared or mixed architectures between vision and language modalities, enabling additional zero-shot capabilities such as visual question answering [21, 22] and image captioning [24]. Our work specifically addresses the adaptation of CLIP families [8, 35], for downstream dense prediction tasks.

Open-vocabulary semantic segmentation. Open-vocabulary semantic segmentation, also known as zero-shot semantic segmentation, aims to segment an image with arbitrary categories described by texts. Recent works have mainly built upon large-scale vision-language models [8,35,43], which could be roughly divided into three types. 1) **Training-free** methods [3, 26, 27, 39, 56] attempt to tap into the inherent localization capabilities of CLIP with minimal modifications. MaskCLIP [56] proposes to extract the value embedding of the last self-attention bock of CLIP's vision encoder for dense prediction tasks. Following this work, many studies [3, 26, 40] generalize the query-key attention to a self-self attention mechanism, such as the value-value attention in CLIPSurgery [26], the query-query and key-key attention in SCLIP [40], and generalized self-self attention combination in GEM [3]. These modifications induce the model to focus more on relevant context, resulting in significantly improved performance. 2) **Unsupervised/Weakly-supervised** methods mainly involve the design of more intricate architectures aimed at explicitly grouping semantic contents with image-only/image-text training samples. GroupViT [44] and SegCLIP [30] introduce the grouping blocks into the vision encoder, whose group tokens serve as class centers for semantic segmentation. OVSegmentor [46] also introduces a set of learnable group tokens via a slot-attention, and performs model training with masked entity completion and cross-image mask consistency proxy tasks. Additionally, PGseg [54] proposes to use both group tokens and prototype tokens to segment the images. TCL [6] and CLIP-S^4 [18] propose to directly generate mask/segment proposals within each image. 3) **Fully-supervised** methods usually involve in-domain fine-tuning, *e.g.*, training on the COCOStuff [4] training set with full dense annotations, and therefore typically achieve better performance compared to training-free and weakly-supervised methods. Existing methods in this category can be broadly categorized into CLIP-based methods [17, 20, 29, 48, 49, 52] and Stable Diffusion-based methods [28, 45, 47].

Our method belongs to training-free open-vocabulary semantic segmentation. We aim to explore the intrinsic localization properties of CLIP from a perspective of feature decomposition.

3 Methodology

In this section, we start by providing an overview of the CLIP model [35] and introducing a baseline for open-vocabulary dense inference in Sec. 3.1. Then, we show how the CLIP baseline fails to achieve satisfactory results which motivates our work in Sec. 3.2. Finally, we elaborate the proposed ClearCLIP for open-vocabulary semantic segmentation in Sec. 3.3.

3.1 Preliminary on CLIP

ViT architecture. A ViT-based CLIP model [35] consists of a series of residual attention blocks. Each of these blocks takes as input a collection of visual tokens $X = [x_{\text{cls}}, x_1, \ldots, x_{h\times w}]^T$, where x_{cls} represents the global class token, and $\{x_i | i = 1, 2, \ldots, h \times w\}$ denote local patch tokens. For brevity, we omit the layer number and format a residual attention block as follows:

$$q = \text{Proj}_q(\text{LN}(X)),\ k = \text{Proj}_k(\text{LN}(X)),\ v = \text{Proj}_v(\text{LN}(X)) \tag{1}$$

$$X_{\text{sum}} = X_{\text{res}} + X_{\text{attn}} = X + \text{Proj}(\text{Attn}_{qk} \cdot v) \tag{2}$$

$$X = X_{\text{sum}} + \text{FFN}(\text{LN}(X_{\text{sum}})), \tag{3}$$

where LN denotes layer normalization, Proj represents a projection layer, and FFN stands for a feed-forward network. X_{res} and X_{attn} denote the residual connection and the attention output. Additionally, $\text{Attn}_{qk} = \text{softmax}(\frac{qk^T}{\sqrt{d_k}})$ represents the q-k attention, where d_k is the dimension of k.

Contrastive pre-training. CLIP employs a transformer-based visual encoder $\mathcal{V}$ and text encoder $\mathcal{T}$ to produce visual representations $X^{\text{visual}}_{\text{cls}}$ and text representations X^{text} for each image-text pair. The pre-training of CLIP is grounded in the contrastive loss. Given a batch of image-text pairs, CLIP is trained to maximize the cosine similarity between the visual representations $X^{\text{visual}}_{\text{cls}}$ and their corresponding text representations X^{text}, while simultaneously minimizing the similarity of these representations from different pairs.

Open-vocabulary dense inference. To adapt CLIP for open-vocabulary semantic segmentation, a baseline approach is to perform dense patch-level classification. Given an image, the image encoder $\mathcal{V}$ is used to extract its visual representations $X^{\text{visual}} = [x^{\text{visual}}_{\text{cls}}, X^{\text{visual}}_{\text{dense}}]^T$, where $X^{\text{visual}}_{\text{dense}} \in \mathbb{R}^{hw\times d}$ denotes the local patch representations in the d-dimensional latent space. For the textual features, object labels with C classes are firstly integrated into a prompt template "`a photo of a {label}.`" to obtain the text descriptions. These descriptions are then fed into CLIP's text encoder to generate the text representations for all C classes $X^{\text{text}} \in \mathbb{R}^{C\times d}$. The final segmentation map $\mathcal{M} \in \mathbb{R}^{hw\times 1}$ is computed as follows:

$$\mathcal{M} = \arg\max_c \cos(X^{\text{visual}}_{\text{dense}}, X^{\text{text}}). \tag{4}$$

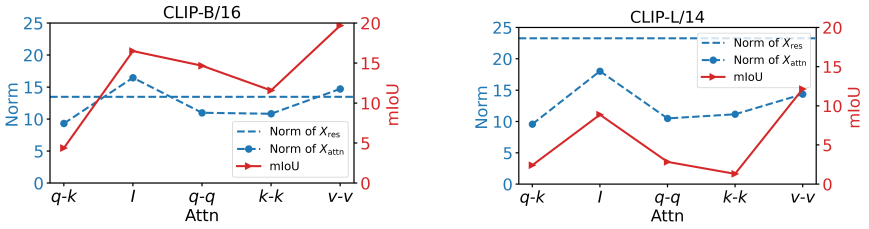


Fig. 2: Comparison of norms and mIoU of different attention mechanisms for CLIP-B/16 (left) and CLIP-L/14 (right). The norm curve of X_{attn} shows a positive correlation with the mIoU curve. A larger norm of X_{res} in CLIP-L/14 impedes the enhancement of performance through the revision of attention mechanisms.

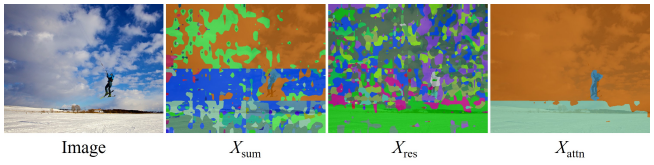
3.2 Motivation

The aforementioned baseline in Eq. (4) often fails to achieve satisfactory results [56]. This is probably because the CLIP is trained with image-level contrastive loss between vision and language, leading to poor alignment between local image regions and text representations [41]. Several studies [3,26,40,56] have attempted to address this challenge with minimal modifications to CLIP without retraining. At the core, they propose to revise the vanilla Attn_{qk} in the last self-attention layer to an identical attention [56] or self-self attention [3, 26, 40], *i.e.*, Attn_{qq}, Attn_{kk} or Attn_{vv}, aiming at re-organizing the spatial information. As shown in Fig. 2, they successfully improve the baseline, with mIoU reaching up to nearly 20.0 from only 4.4 of CLIP with ViT-B/16 architecture (CLIP-B/16) on the COCOStuff dataset. However, there are still several important challenges. Firstly, previous works still generate sub-optimal results with noises in segmentation maps. Secondly, these methods fail to obtain reasonable results when using a larger model, such as ViT-L/14. In Fig. 2, Attn_{qq} and Attn_{kk} are even worse than the vanilla Attn_{qk} with more noises in segmentation maps. Such counter-intuitive phenomena indicates that existing works may have missed some important issues when adapting the CLIP model for dense prediction tasks. In this work, we are curious about *where* and *how* these noises in segmentation results originate and surface.

3.3 ClearCLIP

As explained in Sec. 3.1, a block in the ViT-based CLIP contains three modules, *i.e.*, the residual connection, the self-attention layer and the feed forward network. We delve into these modules to diagnose their effects on open-vocabulary semantic segmentation tasks. Finally, we propose ClearCLIP, a simple yet effective solution to produce clearer and more accurate segmentation maps.

Residual connection. We begin our analysis by comparing the Frobenius norm of the residual connection X_{res} with different attention outputs X_{attn} at the last



Features	mIoU
X_{sum}	4.4
X_{res}	0.01
X_{attn}	11.6

Fig. 3: Open-vocabulary semantic segmentation using different feature maps of CLIP-B/16 model on the COCOStuff dataset. A visualization of an example (left) and quantitative results (right).

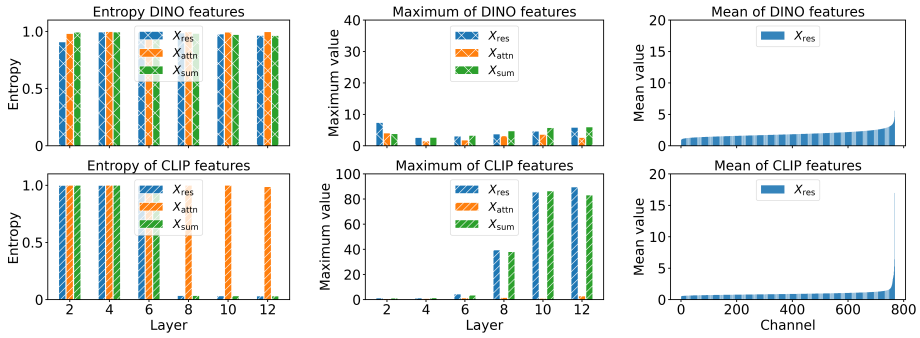
block from CLIP-B/16 and CLIP-L/14 models on the COCOStuff dataset. As illustrated in Fig. 2, we can easily observe the commonalities and distinctions of these two sub-figures. The main commonality is that the mIoU curve and the norm curve of X_{attn} exhibit a certain degree of positive correlation. The distinctions are: 1) the norm of X_{res} in CLIP-B/16 is much smaller than that of CLIP-L/14; and 2) the attention modifications in CLIP-B/16 show consistent improvements over the *q-k* baseline while those in CLIP-L/14 do not. Therefore, we hypothesize that the attention modification is effective only when the influence (or norm) of X_{res} is minimal. In other words, X_{res} substantially impairs the performance of the CLIP family on dense inference tasks.

To investigate this hypothesis, we conduct open-vocabulary semantic segmentation experiments based on CLIP-B/16 using X_{sum}, X_{res} and X_{attn}. Experimental results on the COCOStuff dataset are illustrated in Fig. 3. Surprisingly, we discover that the mIoU of X_{res} is close to zero, suggesting that the residual connection may not be helpful for image segmentation. In contrast, X_{attn} alone could achieve much higher mIoU than X_{sum}. The visualizations in Fig. 3 demonstrate that the noisy segmentation map of CLIP could be decomposed into a muddled map of X_{res} and a clearer map of X_{attn}. According to these experimental results, we can primarily conclude that noises in segmentation maps mainly come from the residual connection.

To gain a deeper understanding of how these noises emerge in semantic segmentation tasks, we conduct a comparative analysis of feature statistics between CLIP-B/16 and DINO-B/16. The latter has demonstrated robust capabilities in learning transferable and semantically consistent dense features for various downstream tasks [16, 23, 31]. We first compare the normalized entropies [15] along layers, which is calculated by

$$H(X^L) = -\frac{1}{\log(hw \times d)} \sum_{i,j} p(X^L_{i,j}) \log p(X^L_{i,j}), \quad p(X^L_{i,j}) = \frac{e^{X^L_{i,j}}}{\sum_{m,n} e^{X^L_{m,n}}}, \quad (5)$$

where X^L denotes the feature map, *i.e.*, X_{sum}, X_{res} and X_{attn}, at the L-th layer of the ViT network. As shown in Fig. 4(a), we can see that the entropy of X^L does not change much across layers for DINO-B/16. On the contrary, for CLIP-B/16, only the entropy of X_{attn} remains the same across the layers, while the entropies of X_{sum} and X_{res} sharply decrease to near-zero. According to Eq. (5),



(a) Entropy **(b)** Maximum **(c)** Mean

Fig. 4: Statistics of three feature maps for DINO-B/16 and CLIP-B/16.

a low entropy indicates that there are a few peak values in X^L. Therefore, we examine the average maximum values of $\max_{i,j} X^L_{i,j}$ in Fig. 4(b). For DINO-B/16, the maximum values of each type of feature maps remain relatively stable along layers, typically lower than 10, resulting in consistent entropies across different layers. In contrast, for CLIP-B/16, the maximum values of X_{res} and X_{sum} gradually increase with the layer depth, peaking nearly 90 times higher at the last layer compared to earlier ones. Consequently, the entropies of X_{res} and X_{sum} sharply decline, approaching near-zero from the middle layers of ViT. Through visualizations of several feature maps (see supplementary material), we empirically found that these peak values appear in a few channels. To verify our observation, we calculate the average normalized mean values of each channel in X_{res} after sorting them in ascending order, and visualize them in Fig. 4(c). We can observe that a few channels dominate the peak values in X_{res} which echoes our discovery from feature maps.

Intuitively, these channel-wise statistics represent the global characteristics of X^L since they are independent of local patterns. If X_{res} and X_{sum} are with low entropy and predominantly influenced by a few channels, it is highly probable that local information is being compromised. As depicted in Eq. (4), distinguishing between two feature vectors with cosine similarity becomes challenging if they share the same dominant channels. While this characteristic is not harmful in itself for image recognition tasks that prioritize global information, it may result in sub-optimal performance when adapting CLIP to dense prediction tasks that emphasize local information. Theoretically, this phenomenon becomes more pronounced in larger vision transformer models with deeper layers. This analysis sheds light on why existing modifications on self-attention fail to yield satisfactory results when applied to the CLIP-L/14 model.

To further demonstrate how X_{res} affects the performance of CLIP, we introduce a scaling factor α[3], $X_{\text{sum}} = X_{\text{res}} + \alpha X_{\text{attn}}$, which controls the relative

[3] SCLIP [40] could be roughly regarded as a special case of $\alpha = 2$, *i.e.*, $\text{Proj}((\text{Attn}_{qq} + \text{Attn}_{kk}) \cdot v) \approx \text{Proj}(2\text{Attn}_{qk} \cdot v) \approx 2X_{\text{attn}}$.

influence of X_{attn} over X_{res}. Our experimental results in Fig. 6 demonstrate that a larger α significantly enhances the performance, which clearly illustrates the adverse impact of X_{res} on the performance. Finally, we propose to directly discard the residual connection to achieve the best performance on dense vision-language inference tasks.

Feed-forward network. The feed-forward network (FFN) in a transformer architecture plays a crucial role in modeling relationships and patterns within the data. However, recent work [14] has revealed that the FFN has a negligible effect on image representation during the inference process. CLIPSurgery [26] finds that the FFN features at the last attention block have a significantly larger cosine angle with the final classification feature, and therefore proposes to discard the FFN for dense prediction tasks. In our work, we empirically find that removing the FFN has minimal effect on open-vocabulary semantic segmentation tasks when applied to the vanilla CLIP model. However, as shown in Fig. 5, when coupled with the removal of the residual connection, discarding the FFN leads to improved results, particularly with a larger model size. The rationale for this improvement is that removing the residual connection significantly alters the input to the FFN, consequently affecting its output. Therefore, removing the FFN output potentially mitigates its negative impact on performance.

Our solution. Based on the above analysis, we propose a straightforward solution to adapt CLIP for open-vocabulary semantic segmentation. Specifically, we propose to use the attention output of the last self-attention layer[4]

$$X^{\text{visual}} = X_{\text{attn}} = \text{Proj}(\text{Attn}_{(\cdot)(\cdot)} \cdot v), \tag{6}$$

for vision-language inference. Inspired by previous works, we could use different combinations of *query-key* in the attention mechanism $\text{Attn}_{(\cdot)(\cdot)}$. In practice, we find that Attn_{qq} consistently achieves better performance in most cases and thus opt to use it by default.

4 Experiments

4.1 Experimental Setups

Datasets & metric. Our solution is extensively evaluated on eight benchmark datasets widely employed for open-vocabulary semantic segmentation. Following [6], these datasets can be categorized into two groups: 1) with background category: PASCAL VOC [13] (**VOC21**), PASCAL Context [34] (**Context60**) and COCO Object [4] (**Object**); and 2) without background category: PASCAL VOC20 [13] (**VOC20**), PASCAL Context59 [34] (**Context59**), COCOStuff [4] (**Stuff**), Cityscapes [11] and ADE20K [55].

[4] The final projection layer is omitted here for brevity.

Table 1: Ablation results based on CLIP-B/16 architecture on five datasets *without* background class. RC denotes the residual connection.

Attn	RC	FFN	VOC20	Context59	Stuff	Cityscapes	ADE20k	Avg.
q-*q*	✓	✓	68.4	24.9	14.7	20.8	7.6	27.3
q-*q*	✓	✗	62.8	25.5	14.6	19.5	6.9	25.9
q-*q*	✗	✓	77.6	31.8	21.0	23.4	14.7	33.7
q-*q*	✗	✗	**80.9**	**35.9**	**23.9**	**30.0**	**16.7**	**37.5**

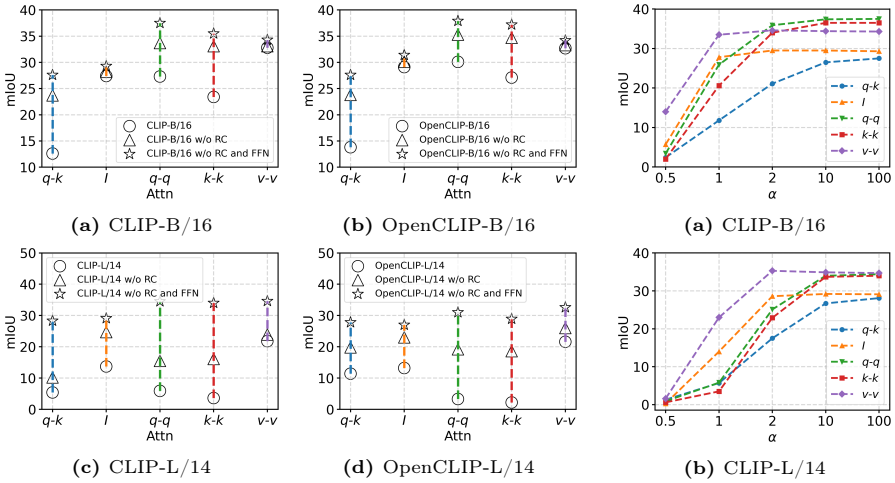
We utilize the implementations provided by MMSegmentation [10], employ a sliding window strategy, and resize input images to have a shorter side of 448 pixels. Following established practices, we avoid text expansions of class names and rely solely on the standard ImageNet prompts [35]. For a fair comparison, no post-processing is applied to any of the methods evaluated. *Our method does not need any retraining or fine-tuning.* Therefore, we can directly evaluate its performance on the validation set of all datasets. For evaluating semantic segmentation tasks, we employ the mean Intersection over Union (mIoU) metric.

Baselines. We compare our method with two types of open-vocabulary semantic segmentation methods: 1) **training-free** methods including CLIP [35], MaskCLIP [56], ReCo [38], CLIPSurgery [26], GEM [3], and SCLIP [40]; and 2) **weakly-supervised** methods including GroupViT [44], SegCLIP [30], OVSegmentor [46], PGSeg [54], ViewCo [36], CoCu [42], and TCL [6]. Unless explicitly mentioned, all reported results are directly cited from the respective papers. Additionally, we include results of the baselines based on CLIP-L/14 using our implementation for comprehensive evaluation.

4.2 Analysis and Discussion

In this section, we present comprehensive experiments to validate the effectiveness of our solution. To ensure a rigorous comparison, our experiments primarily focus on five datasets without the background class.

Ablation study. We conduct ablation studies using the CLIP-B/16 model to assess the effectiveness of our solution. The results are summarized in Tab. 1. Notably, the removal of the residual connection yields a significant performance improvement, increasing the average mIoU from 27.3 to 33.7. This result corroborates our assertion that residual features contain less local information, thereby influencing dense patch prediction. Interestingly, removing the FFN alone does not yield better results. However, the model without both residual connection and FFN together achieves the best performance, with an mIoU of 37.5. This observation is reasonable since removing the residual connection alters the input to the FFN, consequently affecting its output. In this case, removing FFN potentially mitigates the negative impact on performance.



(a) CLIP-B/16 **(b)** OpenCLIP-B/16 **(c)** CLIP-L/14 **(d)** OpenCLIP-L/14

Fig. 5: Ablation study on different architectures and different attention mechanisms.

(a) CLIP-B/16 **(b)** CLIP-L/14

Fig. 6: Segmentation results w.r.t. the scaling factor α.

Different architectures. Given the simplicity of our proposed solution, it could be seamlessly applied to different architectures. We conduct experiments with CLIP [35] and OpenCLIP [8] using ViT-B/16 and ViT-L/14 models. The results regarding the average mIoU on five datasets are depicted in Fig. 5. Our analysis reveals several noteworthy findings: 1) Across different architectures, the consistent achievement of superior segmentation results aligns with the removal of both the residual connection and FFN at the last transformer block. This emphasizes the effectiveness of our solution to adapt vision-language pre-training models for downstream tasks. 2) Notably, the self-self attention consistently outperforms the vanilla q-k attention on our solution. For instance, in the CLIP-B/16 w/o RC and FFN model, the q-q attention yields an average mIoU of 37.5, surpassing the 27.6 mIoU achieved by the q-k attention. 3) For CLIP-L/14 and OpenCLIP-L/14 models, we observe that the self-self attention fails, and the performance of q-q and k-k attentions even falls below that of the vanilla q-k attention. This highlights that existing works aiming at revising the attention mechanism do not address the core problem when adapting CLIP for open vocabulary semantic segmentation. In contrast, our solution of using the attention output leads to significant performance improvements. 4) Interestingly, for models with ViT-B/16 architecture, the improvement of our solution is less pronounced with the identical attention ($\mathbb{I}$) and the v-v attention compared to other attention types. This phenomenon can be attributed to the fact that the $\mathbb{I}$ and v-v attentions tend to sharpen the attention output, thereby increasing the norm of the attention output, as illustrated in Fig. 2. We assert that the enhancement of CLIP-B/16 and OpenCLIP-B/16 under the $\mathbb{I}$ and v-v attentions primarily stems from implicitly eliminating the negative effect of the residual connection. Consequently, explicitly removing the residual connection and FFN

Table 2: Open-vocabulary semantic segmentation quantitative comparison on datasets *without* a background class. † denotes results directly cited from TCL [6]. SCLIP* denotes our reproduced results under the standard setting without class re-name tricks.

Methods	Encoder	VOC20	Context59	Stuff	Cityscape	ADE20k	Avg.
GroupViT† [44]	ViT-S/16	79.7	23.4	15.3	11.1	9.2	27.7
CoCu [42]	ViT-S/16	-	-	13.6	15.0	11.1	-
TCL [6]	ViT-B/16	77.5	30.3	19.6	23.1	14.9	33.1
CLIP [35]	ViT-B/16	41.8	9.2	4.4	5.5	2.1	12.6
MaskCLIP† [56]	ViT-B/16	74.9	26.4	16.4	12.6	9.8	28.0
ReCo† [38]	ViT-B/16	57.7	22.3	14.8	21.1	11.2	25.4
CLIPSurgery [26]	ViT-B/16	-	-	21.9	**31.4**	-	-
SCLIP [40]	ViT-B/16	80.4	34.2	22.4	32.2	16.1	37.1
SCLIP* [40]	ViT-B/16	78.2	33.0	21.1	29.1	14.6	35.2
ClearCLIP	ViT-B/16	**80.9**	**35.9**	**23.9**	30.0	**16.7**	**37.5**
CLIP [35]	ViT-L/14	15.8	4.5	2.4	2.9	1.2	5.4
MaskCLIP [56]	ViT-L/14	30.1	12.6	8.9	10.1	6.9	13.7
SCLIP [40]	ViT-L/14	60.3	20.5	13.1	17.0	7.1	23.6
ClearCLIP	ViT-L/14	80.0	29.6	19.9	27.9	15.0	34.5

leads to limited improvement. However, for models with ViT-L/14 architecture, where the norm of the residual connection is substantially larger, removing both the residual connection and FFN results in significant improvement.

Effect of amplifying the norm of attention output. To further explore the relationship between the residual connection and the attention output in open vocabulary semantic segmentation tasks, we conduct experiments using $\alpha = \{0.1, 1, 2, 10, 100\}$, explicitly amplifying the F-norm of X_{attn} to α^2 times. As shown in Fig. 6, our results reveal a clear trend: as the scaling factor α increases, models with all types of attention exhibit significantly improved performance. As expected, performance sharply declines when α decreases from 1 to 0.5. These findings underscore the importance of enlarging the norm of the attention output to mitigate the negative effects of the residual connection, ultimately leading to substantially improved performance. Hence, our solution of removing the residual connection proves to be simple yet effective. Additionally, these insights help elucidate the superior performance of SCLIP [40], which adopts the q-q plus k-k attention, as this attention mechanism roughly doubles the vanilla attention.

4.3 Comparison to State-of-the-art

Quantitative results. Tab. 2 summarizes the performance of various open-vocabulary semantic segmentation models on datasets without a background class. We observe that our method ClearCLIP achieves the best results on four out of five datasets. ClearCLIP significantly outperforms TCL on all datasets,

Table 3: Open-vocabulary semantic segmentation quantitative comparison on datasets *with* a background class. † denotes results directly cited from TCL [6]. SCLIP* denotes our reproduced results under the standard setting without class re-name tricks.

Methods	Encoder	VOC21	Context60	Object	Avg.
GroupViT† [44]	ViT-S/16	50.4	18.7	27.5	32.2
SegCLIP [30]	ViT-S/16	52.6	24.7	26.5	34.6
OVSegmentor [46]	ViT-B/16	**53.8**	20.4	25.1	33.1
PGSeg [54]	ViT-S/16	53.2	23.8	28.7	35.2
ViewCo [36]	ViT-S/16	52.4	23.0	23.5	33.0
CoCu [42]	ViT-S/16	40.9	21.2	20.3	27.5
TCL [6]	ViT-B/16	51.2	24.3	30.4	35.3
CLIP [35]	ViT-B/16	16.2	7.7	5.5	9.8
MaskCLIP† [56]	ViT-B/16	38.8	23.6	20.6	27.7
ReCo† [38]	ViT-B/16	25.1	19.9	15.7	20.2
CLIPSurgery [26]	ViT-B/16	-	29.3	-	-
GEM [3]	ViT-B/16	46.2	**32.6**	-	-
SCLIP [40]	ViT-B/16	59.1	30.4	30.5	40.0
SCLIP* [40]	ViT-B/16	51.4	30.5	30.0	37.3
ClearCLIP	ViT-B/16	51.8	**32.6**	**33.0**	**39.1**

with an average improvement of 4.4 mIoU. We note that SCLIP also achieves much better performance compared to other methods. This is because SCLIP implicitly attenuates the residual connection by using the q-q plus k-k attention, roughly doubling the attention output. However, our ClearCLIP explicitly removes the residual connection and FFN, resulting in an average improvement of 3.3 mIoU over SCLIP. Interestingly, when adopting the ViT-L/14 model, both MaskCLIP and SCLIP fail to achieve satisfactory results, while our method obtains higher results, at 34.5 mIoU, much better than the 23.6 mIoU of SCLIP. Although this result is not better than those achieved with the ViT-B/16 model, it still demonstrates the better generality of ClearCLIP with different backbones.

We report the results on three datasets with a background class on Tab. 3. The performance of ClearCLIP is significantly better than all weakly-supervised state-of-the-art methods, with an average improvement of 3.8 mIoU over TCL. Additionally, ClearCLIP outperforms SCLIP, with performance improvements of 0.4, 2.1, and 3.0 mIoU on VOC21, Context60, and COCO Object datasets, respectively. These results fully demonstrate the effectiveness of our solution of decomposing CLIP's features for open-vocabulary semantic segmentation.

Qualitative results. In Fig. 7, we present a qualitative comparison between ClearCLIP and three training-free methods, *i.e.*, CLIP, MaskCLIP, and SCLIP. Our observations are summarized as follows: 1) MaskCLIP exhibits good localization ability compared to CLIP but still generates segmentation maps with noticeable noise and many incoherent segments (e.g., those depicting a dog, cat, and duck in the 3rd and 8th columns); 2) SCLIP showcases the capability of

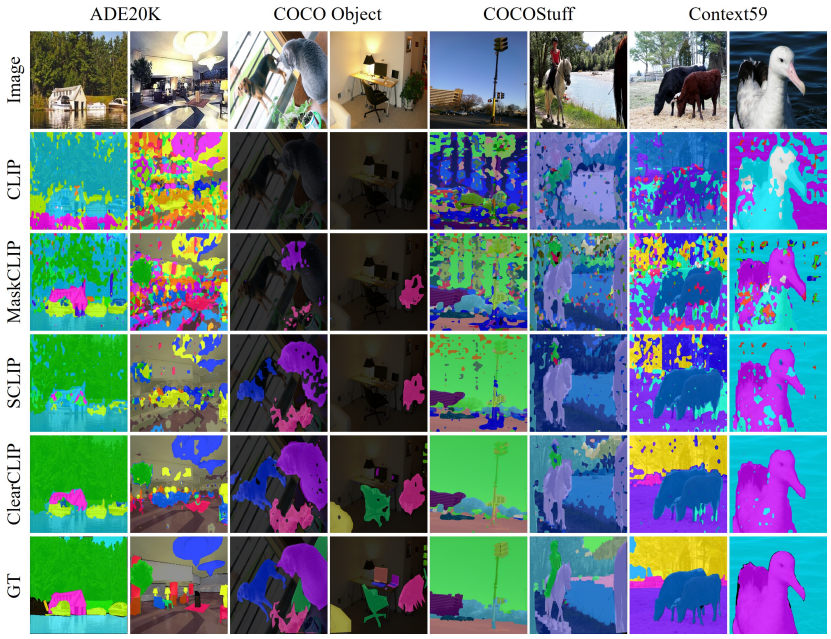


Fig. 7: Qualitative comparison between open-vocabulary segmentation methods.

detecting detailed semantic features with less noise compared to MaskCLIP; 3) ClearCLIP consistently produces much clearer and more accurate fine-grained segmentation maps than the other methods evaluated. These observations validate our attempt to enhance the performance of open-vocabulary semantic segmentation by generating clearer segmentation maps through CLIP's representation decomposition.

5 Conclusion

In this study, we explore the origins and mechanisms of noisy segmentation results when utilizing the CLIP family for open-vocabulary semantic segmentation. We re-examine the architecture of CLIP and conduct a comparative analysis of feature statistics within the residual connection and the attention output. By investigating the differences in norm values across varied sizes of CLIP backbones, we discover that the residual connection serves as the primary source of segmentation noise. Additionally, through a comparative study between CLIP and DINO, we propose that the lack of local information in residual features stems from high-level supervision, which prioritizes global direction. Finally, we introduce ClearCLIP, a simple yet effective solution that removes the residual connection, adopts the self-self attention, and discards the FFN. ClearCLIP demonstrates superior performance and generalizability within the CLIP family.

Acknowledgments. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

1. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. Advances in Neural Information Processing Systems **33**, 25–37 (2020)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision-language transformers. arXiv preprint arXiv:2312.00878 (2023)
4. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
6. Cha, J., Mun, J., Roh, B.: Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11165–11174 (2023)
7. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al.: Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022)
8. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829 (2023)
9. Cho, J., Lei, J., Tan, H., Bansal, M.: Unifying vision-and-language tasks via text generation. In: International Conference on Machine Learning. pp. 1931–1942. PMLR (2021)
10. Contributors, M.: Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark (2020)
11. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
12. Darcet, T., Oquab, M., Mairal, J., Bojanowski, P.: Vision transformers need registers. arXiv preprint arXiv:2309.16588 (2023)
13. Everingham, M., Winn, J.: The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep **2007**(1-45), 5 (2012)
14. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting clip's image representation via text-based decomposition. arXiv preprint arXiv:2310.05916 (2023)

15. Gray, R.M.: Entropy and information theory. Springer Science & Business Media (2011)
16. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. arXiv preprint arXiv:2203.08414 (2022)
17. Han, C., Zhong, Y., Li, D., Han, K., Ma, L.: Open-vocabulary semantic segmentation with decoupled one-pass network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1086–1096 (2023)
18. He, W., Jamonnak, S., Gou, L., Ren, L.: Clip-s4: Language-guided self-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11207–11216 (2023)
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
20. Jiao, S., Wei, Y., Wang, Y., Zhao, Y., Shi, H.: Learning mask-aware clip representations for zero-shot segmentation. arXiv preprint arXiv:2310.00240 (2023)
21. Khan, A.U., Kuehne, H., Gan, C., Lobo, N.D.V., Shah, M.: Weakly supervised grounding for vqa in vision-language transformers. In: European Conference on Computer Vision. pp. 652–670. Springer (2022)
22. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
23. Lan, M., Wang, X., Ke, Y., Xu, J., Feng, L., Zhang, W.: Smooseg: smoothness prior for unsupervised semantic segmentation. Advances in Neural Information Processing Systems **36** (2024)
24. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
25. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
26. Li, Y., Wang, H., Duan, Y., Li, X.: Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:2304.05653 (2023)
27. Li, Y., Li, Z., Zeng, Q., Hou, Q., Cheng, M.M.: Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. arXiv preprint arXiv:2406.00670 (2024)
28. Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Open-vocabulary object segmentation with diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7667–7676 (2023)
29. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
30. Luo, H., Bao, J., Wu, Y., He, X., Li, T.: Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In: International Conference on Machine Learning. pp. 23033–23044. PMLR (2023)
31. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8364–8375 (2022)

32. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020)
33. Mishra, A., Alahari, K., Jawahar, C.: Image retrieval using textual cues. In: Proceedings of the IEEE international conference on computer vision. pp. 3040–3047 (2013)
34. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014)
35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
36. Ren, P., Li, C., Xu, H., Zhu, Y., Wang, G., Liu, J., Chang, X., Liang, X.: Viewco: Discovering text-supervised segmentation masks via multi-view semantic consistency. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=2XLRBjY46O6
37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
38. Shin, G., Xie, W., Albanie, S.: Reco: Retrieve and co-segment for zero-shot transfer. Advances in Neural Information Processing Systems **35**, 33754–33767 (2022)
39. Sun, S., Li, R., Torr, P., Gu, X., Li, S.: Clip as rnn: Segment countless visual concepts without training endeavor. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13171–13182 (2024)
40. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. arXiv preprint arXiv:2312.01597 (2023)
41. Wu, S., Zhang, W., Xu, L., Jin, S., Li, X., Liu, W., Loy, C.C.: Clipself: Vision transformer distills itself for open-vocabulary dense prediction. arXiv preprint arXiv:2310.01403 (2023)
42. Xing, Y., Kang, J., Xiao, A., Nie, J., Shao, L., Lu, S.: Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://openreview.net/forum?id=9iafshF7s3
43. Xu, H., Xie, S., Tan, X.E., Huang, P.Y., Howes, R., Sharma, V., Li, S.W., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying clip data. arXiv preprint arXiv:2309.16671 (2023)
44. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022)
45. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
46. Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., Xie, W.: Learning open-vocabulary semantic segmentation models from natural language supervision. In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2935–2944 (2023)
47. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
48. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
49. Xu, X., Xiong, T., Ding, Z., Tu, Z.: Masqclip for open-vocabulary universal image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 887–898 (2023)
50. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
51. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
52. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. arXiv preprint arXiv:2308.02487 (2023)
53. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
54. Zhang, F., Zhou, T., Li, B., He, H., Ma, C., Zhang, T., Yao, J., Zhang, Y., Wang, Y.: Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation. arXiv preprint arXiv:2310.19001 (2023)
55. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**, 302–321 (2019)
56. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)