

CorrCLIP: Reconstructing Patch Correlations in CLIP for Open-Vocabulary Semantic Segmentation

Dengke Zhang¹

Fagui Liu^{1,2*}

Quan Tang^{2*}

¹South China University of Technology

²Pengcheng Laboratory

csdk@mail.scut.edu.cn, fgliu@scut.edu.cn, tangq@pcl.ac.cn

Abstract

Open-vocabulary semantic segmentation aims to assign semantic labels to each pixel without being constrained by a predefined set of categories. While Contrastive Language-Image Pre-training (CLIP) excels in zero-shot classification, it struggles to align image patches with category embeddings because of its incoherent patch correlations. This study reveals that inter-class correlations are the main reason for impairing CLIP’s segmentation performance. Accordingly, we propose CorrCLIP, which reconstructs the scope and value of patch correlations. Specifically, CorrCLIP leverages the Segment Anything Model (SAM) to define the scope of patch interactions, reducing inter-class correlations. To mitigate the problem that SAM-generated masks may contain patches belonging to different classes, CorrCLIP incorporates self-supervised models to compute coherent similarity values, suppressing the weight of inter-class correlations. Additionally, we introduce two additional branches to strengthen patch features’ spatial details and semantic representation. Finally, we update segmentation maps with SAM-generated masks to improve spatial consistency. Based on the improvement across patch correlations, feature representations, and segmentation maps, CorrCLIP achieves superior performance across eight benchmarks. Codes are available at: <https://github.com/zdk258/CorrCLIP>.

1. Introduction

Open-vocabulary semantic segmentation (OVSS) [4, 55, 69] aims to partition an image into multiple segments and assign a corresponding category to each segment based on category descriptions. Contrastive Language-Image Pre-training (CLIP) [41] model, trained on large-scale image-text pair datasets, shows remarkable zero-shot classification capabilities, providing a viable solution for OVSS.

Applying CLIP to OVSS requires aligning image patches

*Corresponding authors.

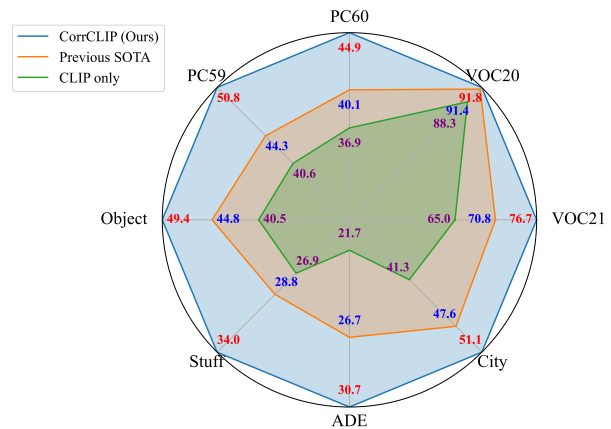


Figure 1. Comparison of OVSS performance on eight benchmarks. “CLIP only” refers to the best performance when only using CLIP.

with corresponding category embeddings. However, due to CLIP’s contrastive image-text training objective, it prioritizes capturing global semantics, resulting in insufficient discriminability between image patches. ClearCLIP [27] demonstrates that removing residual connections and feed-forward network from the last layer of Vision Transformer (ViT) [14] in CLIP’s visual encoder makes patches more distinguishable. Therefore, the key to improving performance lies in the remaining self-attention layer. Self-attention captures correlations between image patches to aggregate contextual information. Although existing methods [28, 32, 51] modify the self-attention to improve CLIP’s segmentation performance, they do so without explicitly identifying the specific patch correlations that impair CLIP’s segmentation capability, limiting further performance gains. To this end, we reveal that the main impediment is inter-class correlations and propose CorrCLIP to reduce the adverse effect of inter-class correlations. As shown in Fig. 1, CorrCLIP significantly expands the performance boundaries of OVSS.

The ideal patch correlations of CLIP for the segmentation task should enhance the distinguishability of patches by ensuring patch features from different classes remain

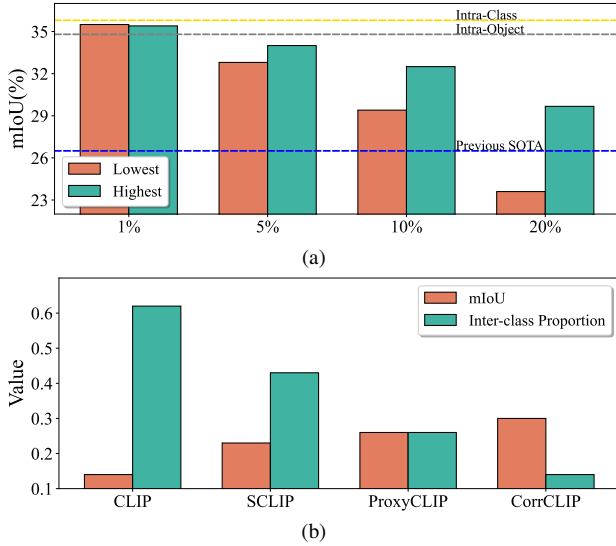


Figure 2. Impact of scope of patch correlations on performance of COCO Stuff. (a) The impact of intra-class and intra-object correlations and different proportions of inter-class correlations on segmentation performance. “Lowest” and “Highest” denote the interactions with the least and most similar inter-class patches, respectively. (b) The inverse relationship between the performance and the proportion of inter-class correlations in current methods.

distinct while also aggregating semantic information effectively to align patches with their corresponding category vectors. We divide patch correlations into two types: intra-class and inter-class correlations. Intuitively, intra-class correlations can achieve the above two points and enhance CLIP’s segmentation capability. However, the impact of inter-class correlations on CLIP’s segmentation capability remains unclear. On the one hand, inter-class correlations may reduce the discriminability of image patches by aggregating information from other categories. On the other, such correlations may facilitate the aggregation of essential semantic information in inter-class image patches, enabling better alignment with corresponding category embeddings.

To investigate the impact of inter-class correlations, we restrict the scope of patch correlations to intra-class. It significantly improves CLIP’s segmentation performance, as shown in Fig. 2a, demonstrating that intra-class correlations benefit CLIP’s segmentation. We then gradually introduce inter-class correlations. Performance declines as the proportion of inter-class correlations increases. Even interactions with the most similar inter-class patches consistently lead to performance degradation in segmentation tasks (we use DINO [6] to measure the similarity, which is more semantically coherent than CLIP [12, 47]). Similar phenomena on other datasets are presented in the supplementary. Additionally, intra-object correlations are sufficient to enhance CLIP’s segmentation capability. This further demonstrates that inter-class correlations are the main reason for

the limited segmentation performance of CLIP.

Existing representative methods [28, 51] modify self-attention to enhance CLIP’s segmentation performance, which could be attributed to the reduction of inter-class correlations. As indicated in Fig. 2b, the performance of these methods decreases as the proportion of inter-class correlations increases. SCLIP [51] employs self-self attention, which reduces inter-class correlations because it makes patches focus on themselves. However, the incoherent CLIP’s features still lead to many inter-class correlations. ProxyCLIP [28] leverages the features of more coherent vision foundation models and applies thresholding to remove low-similarity inter-class correlations. Its performance is still limited because the similarity thresholding cannot reduce high-similarity inter-class correlations.

Therefore, it is necessary to explicitly regulate the scope of patch interactions to more effectively mitigate the adverse effects of inter-class correlations. To this end, we propose scope reconstruction to confine the scope of patch interactions within regions. Specifically, we use SAM [26, 42] to get the region masks of the image and restrict the scope of patch interactions to these regions. As shown in Fig. 2b, our scope reconstruction reduces the proportion of inter-class correlations and significantly improves the segmentation performance (the effectiveness on other datasets is demonstrated in Tab. 4). Because SAM-generated masks may have multiple classes, we employ value reconstruction to obtain more coherent similarity values, thereby further reducing the weight of inter-class correlations.

Apart from patch correlations, we propose feature refinement and map correction from the perspectives of feature representations and segmentation maps. Feature refinement introduces two additional branches to strengthen the patch features’ spatial details and semantic representation. Specifically, the spatial branch incorporates the lower layers’ features into the final patch features to improve spatial details, and the semantic branch utilizes mask class tokens to aggregate global information within region masks to enhance semantic representation. Map correction leverages region masks to update segmentation results, improving spatial consistency. With the careful designs above, CorrCLIP significantly outperforms existing SOTA approaches on eight benchmarks, boosting the averaged mIoU from 48.6% to 53.6%.

Our contributions include: (1) We reveal that inter-class correlations are the primary reason for impairing CLIP’s segmentation performance. Accordingly, we propose scope and value reconstruction to reduce inter-class correlations. (2) We introduce feature refinement to enhance the spatial details and semantic representation of patch features, along with map correction to improve the spatial consistency of segmentation results. (3) We demonstrate the superior performance of CorrCLIP across eight benchmarks.

2. Related Work

Open-Vocabulary Semantic Segmentation. Unlike traditional semantic segmentation [8, 35, 50], OVSS segments an image based on categories described by texts. Recent advancements in OVSS are primarily due to the development of large-scale vision-language models (VLMs) [9, 41, 60]. OVSS methods can be divided into two categories: training-based and training-free. Training-based methods rely on mask annotations [10, 33, 34, 52, 56, 61, 63], images [46, 54], or texts [7, 37, 53, 57, 67]. While training-based methods are typically more effective on specific datasets, they carry the potential risk of reducing the open-vocabulary capacity inherited from VLMs, as described in CaR [49]. In contrast, training-free methods do not need any training and fully leverage the open-vocabulary capabilities of VLMs. Some methods [19, 27, 32, 51, 72] modify the attention mechanism in the final layer from query-key to self-self, enabling patches to focus more on themselves. Others [1, 48, 62] observe that lower-layer attentions exhibit better semantic coherence and thus leverage them to refine the attention maps in the final layer. However, the performance of these approaches is constrained by the inherent limitations of CLIP’s image-text alignment training objective as illustrated in Fig. 1. Consequently, alternative methods [2, 23, 25, 28, 45, 48] utilize characteristics of other foundational models to enhance CLIP’s segmentation capability. Yet, none of the existing methods effectively regulate the scope of patch interactions to mitigate the adverse effects of inter-class correlations. In this paper, we propose scope reconstruction to explicitly define the scope of patch interactions, significantly enhancing CLIP’s segmentation performance.

Vision-Language Pre-training. Vision-language pre-training aims to enable models to learn cross-modal information through weakly supervised training on image-text pairs. This pre-training process allows models to understand the associations between images and corresponding texts. The performance of early research [30, 31, 36] is constrained by the limited size of datasets. Recent studies [22, 41, 64], which leverage large-scale web data, have developed more robust representations. Among these, CLIP [41] stands out as the most popular vision-language model. It employs contrastive learning to align images with corresponding captions, achieving impressive generalization capabilities on unseen data. Subsequent research [9, 17, 18, 38, 59, 60, 65] further enhances CLIP by optimizing training data and processes.

Vision Foundation Models. Vision foundation models (VLMs) undergo pre-training on large-scale datasets to capture general feature representations of the visual world. VLMs learn rich underlying visual patterns, which can be fine-tuned or directly applied to various visual tasks. One type of VLMs is self-supervised models [6, 20, 40, 74],

which aim to learn general-purpose visual features solely from images. Among these, the DINO [6] model shows the power to capture the semantic layout of images [12, 47]. Another type of VLMs is the SAM [26] series, demonstrating impressive zero-shot, class-agnostic segmentation capabilities. Recent advancements have focused on improving the quality of the generated masks [21, 24] and enhancing efficiency to enable broader applications in real-world and mobile scenarios [42, 58, 66, 68, 70, 73]. We capitalize on the strong generalization of SAM and DINO to effectively reduce inter-class correlations.

3. Method

In this section, we first introduce the overall process of adapting CLIP to OVSS in Sec. 3.1. Then, we introduce our scope reconstruction in Sec. 3.2 and value reconstruction in Sec. 3.3. We present feature refinement in Sec. 3.4 and map correction in Sec. 3.5. The overall framework is shown in Fig. 3.

3.1. Preliminary

In ViT, an image is first divided into patches, which are then transformed into token embeddings using a linear layer. These tokens are flattened to form the token sequence represented as $X_C \in \mathbb{R}^{N \times d}$, where N is the sequence length, and d is the dimension of tokens. Next, positional encoding is added to provide position information. The token sequence is then input into the CLIP’s visual encoder. In each preceding layer before the final layer, the token sequence sequentially passes through a multi-head attention network and a feed-forward network, with residual connections applied after each sub-layer.

In the last layer, the token sequence X_C is mapped to query, key, and value embeddings, denoted as Q_C , K_C , and $V_C \in \mathbb{R}^{N \times d}$, respectively. For clarity of exposition, we only show single-head attention. Next, the similarity matrix $S \in \mathbb{R}^{N \times N}$ is computed by the dot product of query embeddings to improve the semantic coherence as illustrated in [27]:

$$S = QQ^T \quad (1)$$

S represents the patch correlations, and its computation is the key to enhancing CLIP’s segmentation capability. In the following subsections, we will introduce how to reconstruct patch correlations to improve segmentation performance. The attention map is derived from the similarity matrix by applying the softmax operation. Then, the attention map is applied to V_C to aggregate information from all patches in the input sequence. To reduce the detrimental impact of noise on the segmentation results, as illustrated in ClearCLIP [27], we remove the residual connections and the feed-forward neural network in the last layer. Subsequently, the sequence of tokens is projected to the final im-

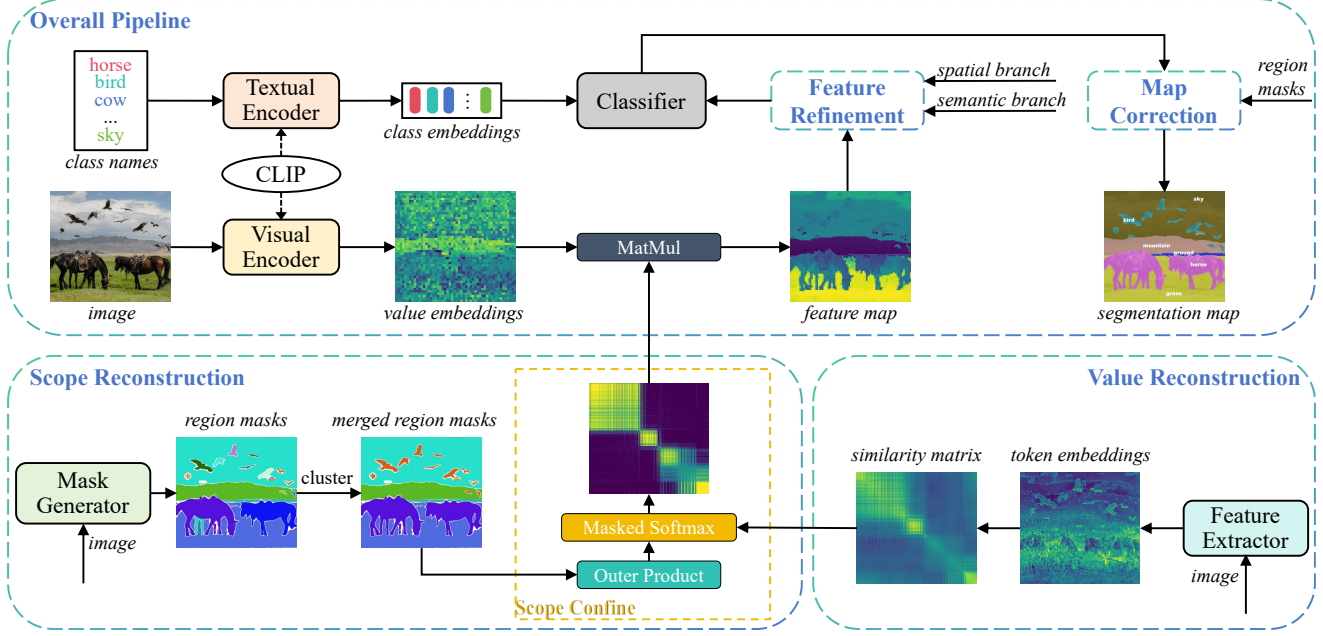


Figure 3. The overall framework of CorrCLIP. CLIP’s visual encoder processes the image to obtain value embeddings, while CLIP’s textual encoder encodes class names to produce class embeddings. A mask generator creates region masks, which are then clustered. A feature extractor derives semantically coherent token embeddings to compute the similarity matrix for value reconstruction. Applying the region masks to the similarity matrix achieves scope reconstruction. Through correlation reconstruction, value embeddings become more distinct. The feature map is refined via spatial and semantic branches, and the refined features are compared with class embeddings to generate the segmentation map, which is further updated using the region masks.

age patch features $F_{img} \in \mathbb{R}^{N \times d}$:

$$Attn = \text{Softmax}\left(\frac{S}{\sqrt{d}}\right) \quad (2)$$

$$F_{img} = \text{Proj}(Attn V_C) \quad (3)$$

Meanwhile, K class names are combined with standard ImageNet prompts [41] to form category descriptions, which are then encoded by CLIP’s textual encoder to generate class embeddings $F_{text} \in \mathbb{R}^{K \times d}$. Then, image patch features are projected to align with the feature space of class embeddings. Finally, the class embeddings serve as the classifier’s parameters to generate the segmentation map:

$$pred = \arg \max_K (\text{Proj}(F_{img}) F_{text}^T) \quad (4)$$

3.2. Scope Reconstruction

Although the self-self attention can enhance the segmentation performance, it is still limited to CLIP’s incoherent patch embeddings, which leads to many inter-class correlations. To address this, we reconstruct the scope of patch interactions using SAM.

SAM uniformly samples points in the image and generates masks for each point. Then thresholding is applied to these masks to discard those with lower confidence and

stability. Subsequently, these masks are downsampled to match the size of the final feature map in CLIP. Consequently, we obtain Z non-overlapping region masks $M = \{m_1, m_2, \dots, m_Z\} \in \mathbb{R}^{Z \times N}$, where $m_i \in \mathbb{R}^N$ represents the i th region mask. The union of all unsegmented regions is denoted by m_0 .

We merge similar region masks to achieve a closer alignment with the true class masks. Specifically, we acquire the features of all segmented regions $F_{region} = \{f_1, f_2, \dots, f_Z\} \in \mathbb{R}^{Z \times d}$ by performing mask average pooling on the token embeddings F_S . Subsequently, we use a clustering algorithm to merge similar regions based on these region features:

$$f_i = \text{Mean}(m_i \odot F_S) \quad (5)$$

$$\hat{M} = \text{Cluster}(M, F_{region}) \quad (6)$$

where $\hat{M} = \{\hat{m}_1, \hat{m}_2, \dots, \hat{m}_z\} \in \mathbb{R}^{z \times N}$ are merged region masks and z is the number of merged region masks. \odot represents the element-wise product. Note that merging masks may result in masks containing multiple classes. However, we believe these classes are highly similar, so the impact of inter-class correlations is minimal, meaning that the benefits of merging masks outweigh the potential drawbacks. We empirically validated this in Tab. 6, observing improved segmentation performance across all datasets.

We then use these merged region masks to calculate the interaction matrix $E \in \mathbb{R}^{N \times N}$ where $E_{i,j} = 1$ means that the i th patch can attend to the j th patch. For patches in unsegmented regions, we set the average value of the similarity matrix as a threshold, allowing only patches with similarity values exceeding this threshold to interact:

$$E = \sum_{i=1}^z \hat{m}_i \otimes \hat{m}_i + (m_0 \otimes m_0) \odot (S > \text{Mean}(S)) \quad (7)$$

where Eq. (7) employs the broadcasting mechanism, and \otimes denotes outer product. Finally, we restrict the scope of patch interactions by applying the masked softmax, where E represents the mask. Eq. (2) is changed as follows:

$$\text{Attn} = \text{Masked Softmax}\left(\frac{S}{\sqrt{d}}, E\right) \quad (8)$$

3.3. Value Reconstruction

Some generated masks may contain multiple classes, causing inter-class correlations. To mitigate the adverse effects of inter-class correlations, we harness DINO’s understanding of semantic layout to construct a more coherent similarity matrix. This could reduce the value of inter-class correlations. Inspired by previous work [32, 44, 51], which combines query and key embeddings in CLIP, we utilize this combination to fully leverage DINO’s capabilities.

Like CLIP, the token sequence in DINO’s final layer is mapped to Q_D , K_D , and $V_D \in \mathbb{R}^{N \times d}$, respectively. In practice, the token sequence lengths of CLIP and DINO may differ due to the variation in patch size, which can be addressed through interpolation. The similarity matrix in Eq. (1) and attention map in Eq. (8) are changed as follows:

$$S = \frac{F_S F_S^T}{\|F_S\|^2} = \frac{(Q_D + K_D)(Q_D + K_D)^T}{\|Q_D + K_D\|^2} \quad (9)$$

$$\text{Attn} = \text{Masked Softmax}\left(\frac{S}{\tau}, E\right) \quad (10)$$

where $\tau < 1$ is a temperature coefficient used to sharpen the attention distribution, amplifying the scores for highly similar patches.

3.4. Feature Refinement

After reconstructing patch correlations, attention is applied to the value embeddings to generate the final patch features. We define this core process as our main branch and propose two additional branches to refine the final patch features.

We present mask class tokens to enhance CLIP’s class token benefits for OVSS. Specifically, we repeat z distinct class tokens at the start of ViT, each corresponding to one mask. In the attention layers, each mask class token exclusively attends to the image tokens within its corresponding mask to capture mask-specific semantics. At

the end of ViT, each image token within a mask is summed with its corresponding mask class token. The final addition enriches the semantic information of the image tokens within each mask, facilitating better alignment with the corresponding text vectors. Additionally, building on recent studies [1, 48, 62] that CLIP’s lower-layer patch features retain richer spatial information, we incorporate these features into the final patch features to enhance spatial details. Finally, image features in Eq. (3) are changed as follows:

$$F_{img} = \text{Proj}(\text{Attn}V_C) + \alpha * \text{Proj}(\text{Attn}V'_C) + \beta * MCT \quad (11)$$

where V'_C is the lower layers’ features and is mapped by the final layer’s value projection inspired by [1]. Further discussion on the spatial branch is presented in the supplementary. MCT represents the mask class tokens. α and β are the coefficients that balance these two additional branches. Empirically, α is set to 1, and β is set to 0.5.

3.5. Map Correction

Spatial consistency is essential in segmentation tasks, referring to maintaining continuity in space. Simply put, it means ensuring that the predictions of adjacent pixels are logically consistent, avoiding abrupt changes such as predicting other objects within a coherent object. Fully supervised methods can learn consistency with mask annotations, but CLIP lacks this capability due to its weak supervision. Although our correlation reconstruction alleviates this issue, it still falls short of the performance achieved with full supervision. To bridge this gap, we reuse the region masks generated above to update the segmentation map. Specifically, we change the categories of all patches within a region to the category that is most common within this region:

$$\text{pred}[m_i] = \text{Mode}(\text{pred}[m_i]), \quad i > 0 \quad (12)$$

4. Experiments

4.1. Dataset and Evaluation Metric

Following prior work, we evaluate our method on the validation sets of five datasets. **Pascal VOC** [16] comprises 1,449 images and serves two benchmarks: VOC21 (21 classes with background) and VOC20 (20 classes without). **Pascal Context** [39] contains 5,104 images and similarly supports PC60 (60 classes) and PC59 (59 classes). **COCO Stuff** [5] (Stuff) includes 5,000 images divided into 171 classes, which encompass both stuff and object categories. **COCO Object** (Object) is a derivative of COCO Stuff, which merges all stuff classes into the background class and has 81 classes. **ADE20k** [71] (ADE) has 2,000 images and 150 classes without the background class. **Cityscapes** [11] (City) has 500 images and 19 classes without the background class. These datasets form eight benchmarks (5 without and 3 with background classes). We compare results using mean Intersection over Union (mIoU).

Method	Size	VOC21	VOC20	PC60	PC59	Object	Stuff	ADE	City	Avg
<i>Training-based</i>										
TCL[7]		55.0	83.2	30.4	33.9	31.6	22.4	17.1	24.0	37.2
CLIP-DINOiser[54]		62.1	80.9	32.4	35.9	34.8	24.6	20.0	31.7	40.3
CoDe[53]	ViT-B/16	57.7	-	30.5	-	32.3	23.9	17.7	28.9	-
CAT-Seg[10]		77.3	94.6	-	57.5	-	-	31.8	-	-
ESC-Net[29]		80.1	97.3	-	59.0	-	-	35.6	-	-
<i>Training-free</i>										
CLIP[41]		11.5	41.9	4.4	9.2	1.6	4.4	2.9	5.0	10.1
MaskCLIP[72]		38.8	74.9	23.6	26.4	20.6	16.4	9.8	12.6	27.9
ClearCLIP[27]		51.8	80.9	32.6	35.9	33.0	23.9	16.7	30.0	38.1
SCLIP[51]		59.1	80.4	30.4	34.2	30.5	22.4	16.1	32.2	38.2
ProxyCLIP[28]		61.3	80.3	35.3	39.1	37.5	26.5	20.2	38.1	42.3
LaVG[23]		62.1	82.5	31.6	34.7	34.2	23.2	15.8	26.2	38.8
CLIPtrase[44]	ViT-B/16	53.0	81.2	30.8	34.9	44.8	24.1	17.0	-	-
NACLIP[19]		64.1	83.0	35.0	38.4	36.2	25.7	19.1	38.3	42.5
Trident[45]		<u>67.1</u>	84.5	<u>38.6</u>	<u>42.2</u>	41.1	<u>28.3</u>	<u>21.9</u>	<u>42.9</u>	<u>45.8</u>
ResCLIP[62]		61.1	86.0	33.5	36.8	35.0	24.7	18.0	35.9	41.4
SC-CLIP[1]		64.6	84.3	36.8	40.1	37.7	26.6	20.1	41.0	43.9
CLIPer[48]		65.9	85.2	37.6	41.7	39.0	27.5	21.4	-	-
CASS[25]		65.8	87.8	36.7	40.2	37.8	26.7	20.4	39.4	44.4
CorrCLIP (Ours)		74.8 (+7.7)	88.8 (+1.0)	44.2 (+5.6)	48.8 (+6.6)	<u>43.7 (-1.1)</u>	31.6 (+3.3)	26.9 (+5.0)	49.4 (+6.5)	51.0 (+5.2)
FreeDA[2]		55.4	87.9	<u>38.3</u>	43.5	37.4	<u>28.8</u>	23.2	36.7	43.9
CaR[49]		67.6	<u>91.4</u>	30.5	39.5	36.6	-	17.7	-	-
ProxyCLIP[28]		60.6	83.2	34.5	37.7	39.2	25.6	22.6	40.1	43.0
ResCLIP[62]	ViT-L/14	54.1	85.5	30.9	34.5	32.5	23.4	18.2	33.7	39.1
SC-CLIP[1]		65.0	88.3	36.9	40.6	40.5	26.9	21.7	<u>41.3</u>	<u>45.2</u>
CLIPer[48]		69.8	90.0	38.0	43.6	43.3	28.7	<u>24.4</u>	-	-
CorrCLIP (Ours)		76.7 (+6.9)	91.5 (+0.1)	44.9 (+6.6)	50.8 (+7.2)	49.4 (+6.1)	34.0 (+5.2)	30.7 (+6.3)	51.1 (+9.8)	53.6 (+8.4)
ProxyCLIP[28]		65.0	83.3	35.4	39.6	38.6	26.8	24.2	42.0	44.4
Trident[45]	ViT-H/14	70.8	88.7	40.1	44.3	42.2	28.6	26.7	47.6	48.6
CorrCLIP (Ours)		76.4 (+5.6)	91.8 (+3.1)	42.5 (+2.4)	47.9 (+3.6)	48.4 (+6.2)	32.7 (+4.1)	28.8 (+2.1)	49.9 (+2.3)	52.3 (+3.7)

Table 1. Comparison with state-of-the-art OVSS methods on eight benchmarks in three different sizes of CLIP. Bold fonts indicate the optimal methods and underlined fonts indicate the suboptimal methods. ‘‘Avg’’ represents the averaged mIoU across eight benchmarks.

4.2. Implementation Details

Our model has three different sizes based on CLIP: ViT-B/16 [60], ViT-L/14 [60], and ViT-H/14 [9]. All other configurations for the three variants are identical. The backbone of DINO is ViT-B/8 [6]. We use SAM2 [42] with MAE [20] pre-trained Hiera-L [3, 43]. Results for other mask generators are presented in the supplementary. We collect 32×32 prompt points in a grid manner to generate region masks of an image. ‘‘pred_iou_thresh’’ and ‘‘stability_score_thresh’’ in mask thresholding are set to 0.7 for all datasets. The clustering algorithm used in mask merging is DBSCAN [15], where the neighborhood radius is set to 0.2, and the minimum number of samples is set to 1 for all datasets. The temperature coefficient τ is set to 0.25.

We resize the images to meet the varying specifications of different datasets: a shorter side of 336 pixels for Pascal VOC, Pascal Context, and COCO, and 448 pixels for Cityscapes and ADE20K. We perform slide inference with a 336×336 window and 112×112 stride.

4.3. Comparison with State-of-the-Art Methods

We compare CorrCLIP with state-of-the-art training-free OVSS methods across eight benchmarks. In these methods,

MaskCLIP [72], CaR [49], SCLIP [51], ClearCLIP [27], CLIPtrase [44], NACLIP [19], SC-CLIP[1], and ResCLIP [62] only use CLIP. The incoherent patch features of CLIP constrain their performance. LaVG [23], ProxyCLIP [28], Trident [45], CLIPer [48], CASS [25], and FreeDA [2] use foundation models to alleviate CLIP’s limitation. However, they do not effectively confine the scope of correlations.

The quantitative results are summarized in Tab. 1. All three variants of our CorrCLIP demonstrate superior performance across eight benchmarks, significantly outperforming other methods. In particular, CorrCLIP achieves increases of 5.2%, 8.4%, and 3.7% in averaged mIoU across three variants. We also compare our approach with several state-of-the-art training-based methods [7, 10, 29, 53, 54]. CorrCLIP surpasses weakly-supervised methods and narrows the gap to fully-supervised ones. It further demonstrates superior generalization by outperforming fully-supervised models on out-of-distribution datasets, as detailed in the supplementary.

In Fig. 4, we provide qualitative comparisons of our method, CorrCLIP, with ClearCLIP [27], ProxyCLIP [28], SC-CLIP [1], and Trident [45]. Due to the explicitly restricted interaction scope, our method correctly identified

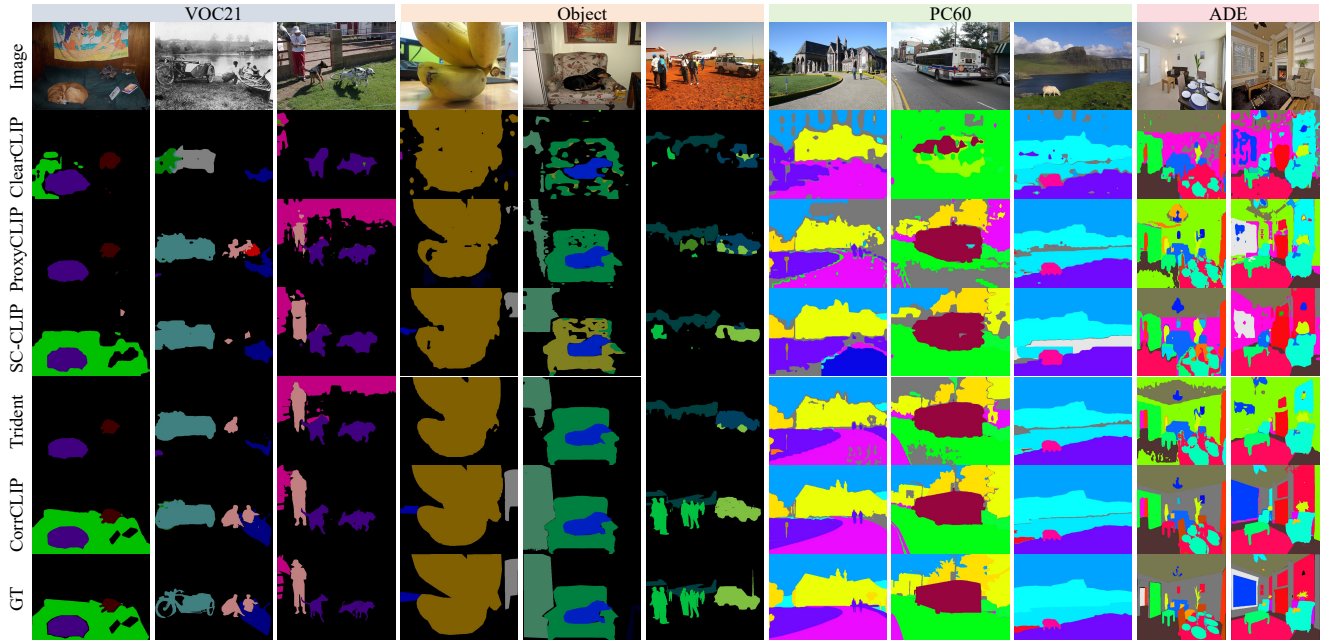


Figure 4. Qualitative comparison between our method CorrCLIP and the other four methods. “GT” denotes ground truth.

SR	VR	MC	FR	VOC21	PC60	Object	ADE	City
				53.6	31.9	32.0	19.3	30.5
✓				68.1	39.6	41.5	24.0	42.0
✓	✓			68.5	40.3	42.0	24.0	43.2
✓	✓	✓		72.5	42.0	43.7	25.3	48.3
✓	✓	✓	✓	74.8	44.2	43.7	26.9	49.4

Table 2. Impact of integrating different components.

objects misclassified by other approaches. Our method also demonstrates strong object continuity and significantly reduces noise compared to other methods. Additional qualitative comparisons are provided in the supplementary.

4.4. Ablation Study

We perform a series of ablation studies to investigate the effects of each component in our method.

Impact of Integrating Different Components. We investigate the impact of four components of CorrCLIP outlined in Tab. 2. These components are scope reconstruction (SR), value reconstruction (VR), map correction (MC), and feature refinement (FR). We use ClearCLIP [27] as the baseline, which substitutes query-key with query-query and removes the feed-forward network and residual connections. SR significantly enhances CLIP’s segmentation performance, yielding improvements of 14.5%, 7.7%, 9.5%, 4.7%, and 11.5% mIoU on VOC, PC, Object, ADE, and City, respectively. VR has a further improvement. Incorporating MC and FR also leads to performance gains. These results demonstrate the effectiveness of each component.

Method	VOC21	PC60	Object	ADE	City
SCLIP[51]	59.6	31.7	33.5	16.5	32.3
+SR	66.5 _{+6.9}	35.9 _{+4.2}	37.0 _{+3.5}	19.8 _{+3.3}	39.0 _{+6.7}
ProxyCLIP[28]	61.3	35.3	37.5	20.2	38.1
+SR	67.0 _{+5.7}	38.2 _{+3.1}	41.7 _{+4.2}	21.6 _{+1.4}	40.8 _{+2.7}
SC-CLIP[1]	64.6	36.8	37.7	20.1	41.0
+SR	66.0 _{+1.4}	37.9 _{+1.1}	38.8 _{+1.1}	20.9 _{+0.8}	42.4 _{+1.4}

Table 3. Effectiveness of Scope Reconstruction on other methods.

The computational costs of each component and the corresponding analyses are presented in the supplementary.

Effectiveness of Scope Reconstruction. Our proposed scope reconstruction can be seamlessly integrated into other methods [1, 28, 51]. As shown in Tab. 3, our SR can further enhance the performance of these methods, as they do not explicitly confine the scope of correlations. Additionally, a key limitation of the SR process is the required down-sampling of generated masks to match the low-resolution patch feature map, which can introduce quantization errors. We hypothesize that mitigating this error by increasing the mask resolution would improve performance. To validate this, we upsample the patch feature map via interpolation to generate larger final masks. As shown in Tab. 4, the results confirm our hypothesis: SR performance steadily improves with mask size, underscoring the benefit of reducing down-sampling errors. However, since upsampling is highly time- and memory-intensive, we set the final mask size to 42×42 . **Ablation Study on Value Reconstruction.** We explore the effect of different features used to compute the similarity

Mask Size	VOC21	PC60	Object	ADE	City
No SR	53.6	31.9	32.0	19.3	30.5
21×21	64.7	37.8	39.5	22.8	38.4
42×42	68.1	39.6	41.5	24.0	42.0
63×63	69.3	40.1	42.3	24.4	43.3
84×84	70.2	40.7	43.0	24.8	44.7

Table 4. Impact of the final masks’ size on scope reconstruction.

Sim	Model	VOC21	PC60	Object	ADE	City
Unif	-	73.4	43.4	43.8	26.6	47.0
Q-K	CLIP-B	69.5	41.6	42.4	25.7	45.1
Q-Q	CLIP-B	73.8	43.4	44.3	26.7	48.0
X-X	DINO-B	73.4	43.8	43.1	26.9	49.0
QK-QK	DINO-S	74.3	43.9	43.7	26.8	48.3
QK-QK	DINO-B	74.2	44.0	43.4	26.8	49.1

Table 5. Effect of different features used to compute the similarity values after scope reconstruction. “Unif” denotes all patches have the same similarity. “X” denotes output features.

Cluster	VOC21	PC60	Object	ADE	City
×	74.2	44.0	43.4	26.8	49.1
✓	74.8	44.2	43.7	26.9	49.4

Table 6. Effectiveness of mask merging.

values after SR. As shown in Tab. 5, our proposed combination of DINO’s query and key embeddings yields the best average performance. Note that features from smaller DINO (ViT-S) also show comparable performance. Interestingly, we found that even without computing similarity (i.e., uniform similarity across all patches), our method still achieves strong performance. Meanwhile, CLIP’s original query-key computation mechanism yields inferior results due to its generation of highly incoherent similarity values. However, it still outperforms existing SOTA methods, further validating the effectiveness of our proposed SR.

Effectiveness of Mask Merging. We explore the effectiveness of mask merging on multiple benchmarks. As shown in Tab. 6, mask merging boosts segmentation performance across all datasets, as enhanced intra-class correlations outweigh the adverse effects of inter-class correlations. This is reasonable because mask merging generally introduces high-similarity inter-class correlations, which have less adverse effects on CLIP’s segmentation performance.

Ablation Study on CLIP. We investigate the effect of different types and sizes of CLIP models on CorrCLIP. We focus on four specific types of CLIP: CLIP [41], OpenCLIP [9], MetaCLIP [60], and DFNCLIP [17]. These models share the same architecture and only differ from training data and processes. As shown in Tab. 7, the segmentation performance of the original CLIP and ClearCLIP (our adopted baseline) does not improve with the enhancement of CLIP’s zero-shot classification capability. In contrast, our method’s segmentation performance is positively corre-

Type	Size	Acc	Plain	ClearCLIP	CorrCLIP
CLIP		68.3	12.7	37.8	47.0
OpenCLIP	ViT-B	70.2	13.9	38.6	48.1
MetaCLIP		72.1	11.0	37.7	49.1
DFNCLIP		76.2	12.6	39.4	49.5
CLIP		75.5	5.3	35.2	49.0
OpenCLIP	ViT-L	75.3	11.3	32.1	48.7
MetaCLIP		79.2	4.6	35.1	51.6
DFNCLIP		81.4	2.6	34.6	50.7
OpenCLIP		78.0	7.5	35.5	50.2
MetaCLIP	ViT-H	80.5	3.2	24.0	50.2
DFNCLIP		83.4	1.3	34.4	50.6

Table 7. Impact of varied CLIP on the performance of different OVSS methods. OVSS performance is evaluated on five benchmarks without the background class. “Acc” is the zero-shot accuracy on ImageNet [13].

Semantic	Spatial	VOC21	PC60	Object	ADE	City
		72.5	42.0	43.7	25.3	48.3
✓		74.1	43.2	44.0	26.5	47.6
	✓	72.9	43.3	42.6	26.4	49.9
✓	✓	74.8	44.2	43.7	26.9	49.4

Table 8. Effectiveness of the spatial branch and semantic branch.

lated with CLIP’s zero-shot capability, demonstrating that our approach fully exploits the potential of CLIP.

Ablation Study on Additional Branches. We investigate the roles of our proposed two additional branches, as shown in Tab. 8. The semantic and spatial branches improve segmentation performance on most datasets. However, their individual application leads to performance degradation on the City and Object datasets, as modifying the final patch features may disrupt the alignment between patches and category embeddings. Combining the two branches synergistically enhances their effects, guaranteeing consistent performance improvements across all evaluated datasets.

5. Conclusion

In this paper, we reveal that inter-class correlations significantly impair CLIP’s segmentation performance. Accordingly, we propose scope reconstruction to confine the scope of patch interactions, reducing inter-class correlations. To alleviate the problem that scope reconstruction cannot completely eliminate inter-class correlations, we present value reconstruction to compute more coherent similarity values, reducing the weight of inter-class correlations. Moreover, we introduce feature refinement to enhance the spatial granularity and semantic richness of patch features and map correction to improve the spatial consistency of segmentation maps. Armed with the above designs, our method significantly enhances CLIP’s segmentation capability and demonstrates superior performance across eight benchmarks.

6. Acknowledgment

This work was partially supported by the Guangdong Major Project of Basic and Applied Basic Research (2019B030302002), the Science and Technology Project of Guangdong Province (2021B1111600001), the Major Key Project of PCL (PCL2025AS208, PCL2025AS213), and the National Natural Science Foundation of China (U24B20151).

References

- [1] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*, 2024. 3, 5, 6, 7
- [2] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *CVPR*, pages 3689–3698, 2024. 3, 6
- [3] Daniel Bolya, Chaitanya Ryali, Judy Hoffman, and Christoph Feichtenhofer. Window attention is bugged: How not to interpolate position embeddings. In *ICLR*, 2023. 6
- [4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NeurIPS*, 2019. 1
- [5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 5
- [6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 2, 3, 6
- [7] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *CVPR*, pages 11165–11174, 2023. 3, 6
- [8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 3
- [9] 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 *CVPR*, pages 2818–2829, 2023. 3, 6, 8
- [10] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024. 3, 6
- [11] 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 *CVPR*, pages 3213–3223, 2016. 5
- [12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 2, 3
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 8
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1
- [15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 6
- [16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 5
- [17] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 3, 8
- [18] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023. 3
- [19] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, pages 5061–5071, 2025. 3, 6
- [20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3, 6
- [21] You Huang, Wenbin Lai, Jiayi Ji, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. Hrsam: Efficiently segment anything in high-resolution images. *arXiv preprint arXiv:2407.02109*, 2024. 3
- [22] 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 *ICML*, pages 4904–4916, 2021. 3
- [23] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *ECCV*, pages 143–164. Springer, 2024. 3, 6
- [24] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. In *NeurIPS*, 2024. 3
- [25] Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *CVPR*, pages 15033–15042, 2025. 3, 6
- [26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2, 3

- [27] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, 2024. 1, 3, 6, 7
- [28] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, 2024. 1, 2, 3, 6, 7
- [29] Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *CVPR*, pages 26081–26090, 2025. 6
- [30] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 3
- [31] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137, 2020. 3
- [32] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 1, 3, 5
- [33] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 3
- [34] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *CVPR*, pages 3491–3500, 2024. 3
- [35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 3
- [36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3
- [37] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, pages 23033–23044, 2023. 3
- [38] Jiawei Ma, Po-Yao Huang, Saining Xie, Shang-Wen Li, Luke Zettlemoyer, Shih-Fu Chang, Wen-Tau Yih, and Hu Xu. Mode: Clip data experts via clustering. In *CVPR*, pages 26354–26363, 2024. 3
- [39] 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 *CVPR*, pages 891–898, 2014. 5
- [40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 3, 4, 6, 8
- [42] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chaoyuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 3, 6
- [43] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, pages 29441–29454, 2023. 6
- [44] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *ECCV*, 2024. 5, 6
- [45] Yuheng Shi, Mingjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. *arXiv preprint arXiv:2411.09219*, 2024. 3, 6
- [46] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*, pages 33754–33767, 2022. 3
- [47] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021. 2, 3
- [48] Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2411.13836*, 2024. 3, 5, 6
- [49] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *CVPR*, pages 13171–13182, 2024. 3, 6
- [50] Quan Tang, Chuanjian Liu, Fagui Liu, Jun Jiang, Bowen Zhang, CL Philip Chen, Kai Han, and Yunhe Wang. Rethinking feature reconstruction via category prototype in semantic segmentation. *IEEE TIP*, 2025. 3
- [51] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, 2024. 1, 2, 3, 5, 6, 7
- [52] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, et al. Use: Universal segment embeddings for open-vocabulary image segmentation. In *CVPR*, pages 4187–4196, 2024. 3
- [53] Ji-Jia Wu, Andy Chia-Hao Chang, Chieh-Yu Chuang, Chun-Pei Chen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Yung-Yu Chuang, and Yen-Yu Lin. Image-text co-decomposition for text-supervised semantic segmentation. In *CVPR*, pages 26794–26803, 2024. 3, 6
- [54] Monika Wysockańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciański, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks. In *ECCV*, 2024. 3, 6

- [55] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, pages 8256–8265, 2019. 1
- [56] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *CVPR*, pages 3426–3436, 2024. 3
- [57] Yun Xing, Jian Kang, Aoran Xiao, Jiahao Nie, Ling Shao, and Shijian Lu. Rewrite caption semantics: Bridging semantic gaps for language-supervised semantic segmentation. In *NeurIPS*, 2023. 3
- [58] Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, et al. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In *CVPR*, pages 16111–16121, 2024. 3
- [59] Hu Xu, Po-Yao Huang, Xiaoqing Ellen Tan, Ching-Feng Yeh, Jacob Kahn, Christine Jou, Gargi Ghosh, Omer Levy, Luke Zettlemoyer, Wen-tau Yih, et al. Altogether: Image captioning via re-aligning alt-text. *arXiv preprint arXiv:2410.17251*, 2024. 3
- [60] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. In *ICLR*, 2024. 3, 6, 8
- [61] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 3
- [62] Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. In *CVPR*, pages 29968–29978, 2025. 3, 5, 6
- [63] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, pages 32215–32234, 2023. 3
- [64] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 3
- [65] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *CVPR*, pages 11975–11986, 2023. 3
- [66] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 3
- [67] Fei Zhang, Tianfei Zhou, Boyang Li, Hao He, Chaofan Ma, Tianjiao Zhang, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Uncovering prototypical knowledge for weakly open-vocabulary semantic segmentation. In *NeurIPS*, 2023. 3
- [68] Zhuoyang Zhang, Han Cai, and Song Han. Efficientvit-sam: Accelerated segment anything model without performance loss. *arXiv preprint arXiv:2402.05008*, 2024. 3
- [69] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *ICCV*, pages 2002–2010, 2017. 1
- [70] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 3
- [71] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 5
- [72] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712, 2022. 3, 6
- [73] Chong Zhou, Xiangtai Li, Chen Change Loy, and Bo Dai. Edgesam: Prompt-in-the-loop distillation for on-device deployment of sam. *arXiv preprint arXiv:2312.06660*, 2023. 3
- [74] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022. 3