DeCLIP: Decoupled Learning for Open-Vocabulary Dense Perception

Junjie Wang[1] Bin Chen[2,3,*] Yulin Li[1] Bin Kang[3] Yichi Chen[3] Zhuotao Tian[1,*]
[1]School of Computer Science and Technology, HIT, Shenzhen
[2]International Research Institute for Artificial Intelligence, HIT, Shenzhen
[3]University of Chinese Academy of Sciences

Abstract

Dense visual prediction tasks have been constrained by their reliance on predefined categories, limiting their applicability in real-world scenarios where visual concepts are unbounded. While Vision-Language Models (VLMs) like CLIP have shown promise in open-vocabulary tasks, their direct application to dense prediction often leads to suboptimal performance due to limitations in local feature representation. In this work, we present our observation that CLIP's image tokens struggle to effectively aggregate information from spatially or semantically related regions, resulting in features that lack local discriminability and spatial consistency. To address this issue, we propose DeCLIP, a novel framework that enhances CLIP by decoupling the self-attention module to obtain "content" and "context" features respectively. The "content" features are aligned with image crop representations to improve local discriminability, while "context" features learn to retain the spatial correlations under the guidance of vision foundation models, such as DINO. Extensive experiments demonstrate that DeCLIP significantly outperforms existing methods across multiple open-vocabulary dense prediction tasks, including object detection and semantic segmentation. Code is available at https://github.com/xiaomoguhz/DeCLIP.

1. Introduction

In the era of deep learning, dense prediction tasks like object detection [30, 38] and image segmentation [8, 39] have rapidly advanced and are widely used. However, traditional methods [4, 26, 66] recognize only a fixed set of predefined categories. This restriction hinders the practical application of these methods in real-world settings, where the range of visual concepts is virtually boundless. Consequently, increasing attention has been drawn to open-vocabulary methods [10, 47, 50, 61], which aim to detect and segment objects from any category using textual descriptions.

*Corresponding authors

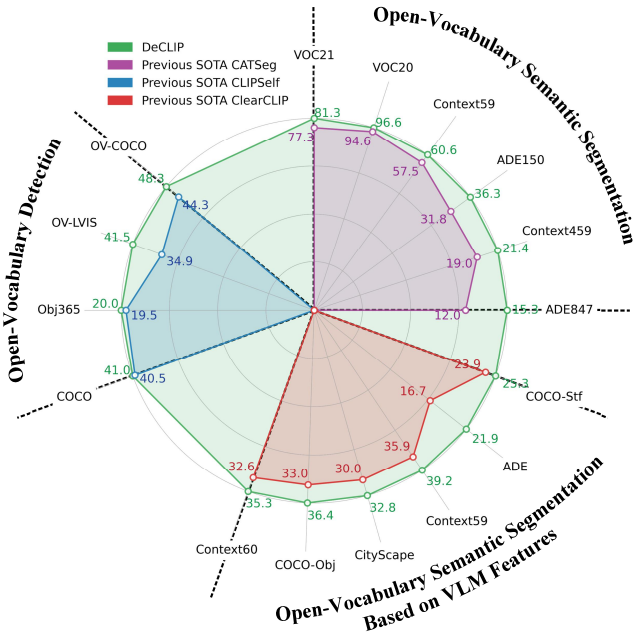

Figure 1. DeCLIP outperforms previous state-of-the-art models on a broad range of open-vocabulary dense prediction benchmarks.

Building on the success of Vision-Language Models (VLMs) [9, 27, 36, 43] pre-trained on image-text pairs, such as CLIP [36], researchers have started leveraging these models for open-vocabulary dense prediction tasks. Among these [5, 46–49, 62], transfer-learning approaches [7, 17, 23, 46, 48, 58] have shown outstanding performance. These methods utilize the image encoder of VLM as a feature extractor and exclusively train lightweight task-specific components. Whereas using VLMs as feature extractors offers significant advantages due to their comprehensive pre-training, directly applying these image-level models to dense prediction often leads to domain shift issues [48, 50].

What hinders CLIP in dense perception? To assess VLM's constraints in dense perception, we analyze CLIP's attention maps across various layers (Figure 3(a)). Our experiments reveal that CLIP's [CLS] token may interfere

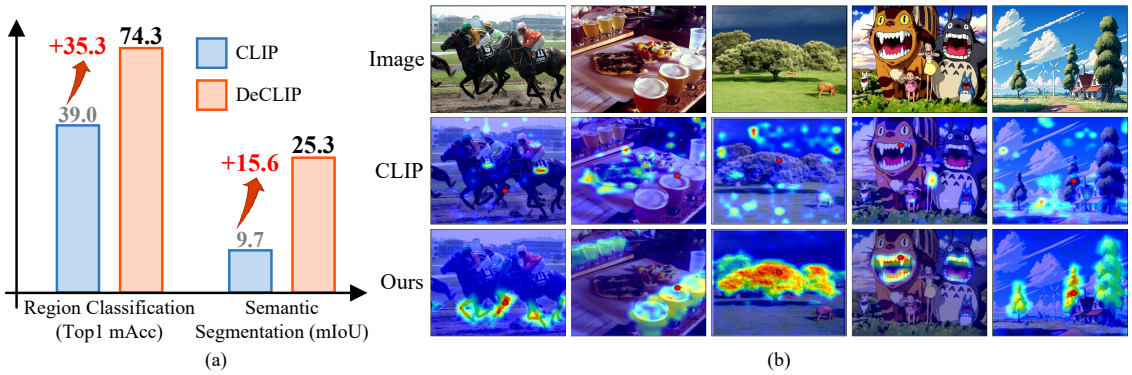

Figure 2. **Quantitative and qualitative comparisons between our method and CLIP. (a)** Performance comparisons of open-vocabulary dense predictions on COCO [29]. **(b)** Attention map comparisons, with the anchor image token marked in red.

with the correlations among other image tokens, leading to suboptimal performance in dense prediction tasks.

Specifically, we have observed that in deeper layers (behind the 9th layer), the [CLS] token shifts focus away from primary objects within the image and attends highly to certain background tokens, as highlighted by the bright spots in the first row of Figure 3(a). Moreover, image tokens (rows 2 and 3, Figure 3(a)) exhibit similar behavior with the [CLS] token, showing high attention to certain background tokens regardless of their positions.

This observation sheds light on why CLIP struggles in dense prediction tasks: its image tokens fail to aggregate information from spatially or semantically related regions, resulting in dense features that lack local discriminability and spatial consistency. As shown in Figure 2(a), directly using CLIP features on the COCO dataset yields relatively inferior performance in open-vocabulary region classification and semantic segmentation. To tackle this, an intuitive approach is to enhance CLIP's local representations through fine-tuning. However, balancing the optimizations of both local feature spatial correlations and vision-language semantic alignment within a unified architecture becomes a new challenge. Therefore, *is it feasible to disentangle CLIP's features and apply separate guiding constraints to obtain diverse features within a unified architecture?*

Our solution. To address these challenges, we propose DeCLIP, a general unsupervised fine-tuning method aimed at enhancing both the discriminability and spatial consistency of CLIP's local features. The core idea is to decouple the self-attention module of CLIP and learn from different teacher models separately.

Specifically, DeCLIP decouples the features in the self-attention module into "*content*" and "*context*" components. The "content" features, responsible for local discriminability, are fine-tuned by aligning pooled region features with their corresponding image crop [CLS] representations. Meanwhile, the "context" features, responsible for spatial consistency, are learned from the feature correlations generated by Vision Foundation Models (VFMs). This decoupled distillation design effectively mitigates optimization conflicts, improving the generalization ability when applying CLIP to downstream open-vocabulary dense prediction tasks. As shown in Figure 2, DeCLIP significantly outperforms CLIP in local discriminability and spatial consistency. To summarize, our contributions are as follows:

- We analyze CLIP and find that its limitation in open-vocabulary dense prediction arises from image tokens failing to aggregate information from spatially or semantically related regions.
- To address this issue, we propose DeCLIP, a simple yet effective unsupervised fine-tuning framework, to enhance the discriminability and spatial consistency of CLIP's local features via a decoupled feature enhancement strategy.
- Extensive experiments demonstrate that DeCLIP can be decently applied to mainstream open-vocabulary dense prediction tasks, including object detection and semantic segmentation. As illustrated in Figure 1, DeCLIP outperforms state-of-the-art methods across a broad range of benchmarks, achieving superior performance metrics in all evaluated task domains.

2. Background and Motivation

In the following, we provide a concise overview of foundational concepts pertinent to this study in Section 2.1, and highlight important findings in Section 2.2, which offer valuable insights for motivating the proposed approach.

2.1. Preliminaries

Contrastive Language-Image Pre-training (CLIP) [36] is built upon two encoders, one for images and one for text. The visual encoder of CLIP can be a CNN series [15, 31] or ViT [12], and the text encoder is a Transformer [44]. This paper focuses on the CLIP model with the ViT architecture,

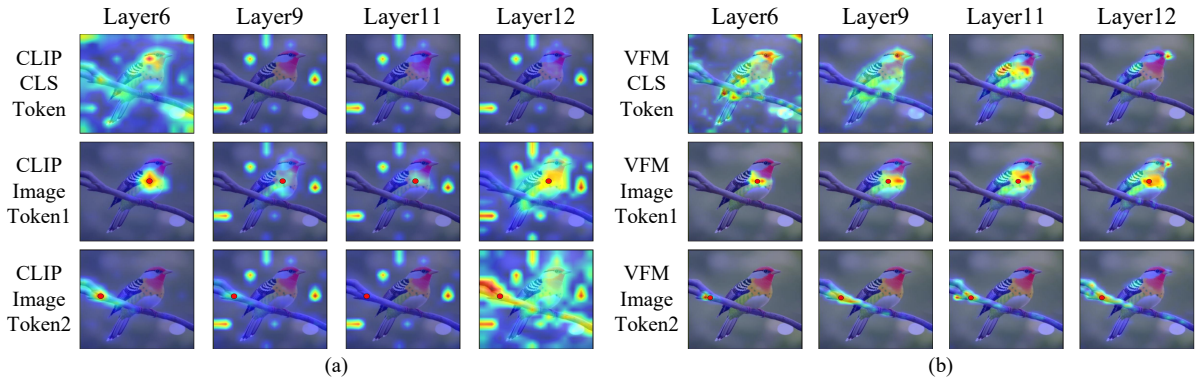

Figure 3. **Visualization of attention maps across different encoding layers of CLIP and VFM.** The attention weights are calculated at a low resolution, then averaged across different heads, and finally upsampled to the original image resolution for visualization. The anchor image token is marked in red. We observe the occurrence of the "proxy" token phenomenon in CLIP, but not in VFM. Furthermore, when the position of the anchor image token is shifted, VFM shows a better correlation for image tokens with the same semantics.

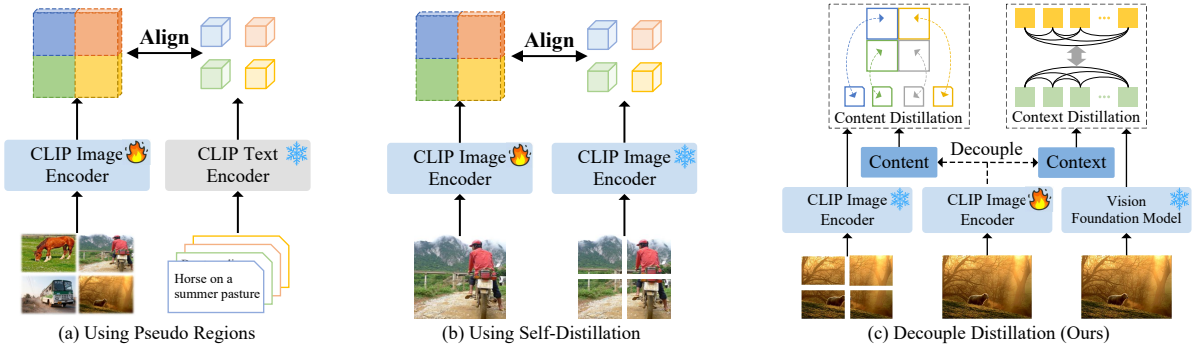

Figure 4. **Pre-fine-tuning methods for adapting CLIP to dense prediction tasks.** Existing work considers establishing region-text alignment through cost-effective methods via: **(a)** using images as pseudo regions or **(b)** using self-distillation on image patches. The former regards the entire image as a region, which results in a loss of details. The latter uses self-distillation on the image patches thereby gaining more fine-grained information, but still fails to apply to pixel-level image segmentation. **(c)** Unlike prior approaches, we use VFM to guide the spatial consistency of CLIP's features, and decouple CLIP's features for distillation separately to avoid optimization conflicts.

which adopts the [CLS] token to represent the overall features of an image. CLIP learns vision-language alignment by maximizing the cosine similarity between the [CLS] token and text features of matched image-text pairs, and minimizing the similarity for unmatched pairs.

Dense feature extraction with CLIP. ViT-based CLIP consists of a series of stacked attention blocks. For example, the ViT-B version of CLIP includes 12 attention block layers. Let $\mathbf{X} = \{\boldsymbol{x}_0, \boldsymbol{x}_1, \cdots, \boldsymbol{x}_{h\times w}\}$ denotes the input to the last attention block, where $\boldsymbol{x}_i \in \mathbb{R}^{1\times D}$. The computation within this attention block can be expressed as:

$$\mathbf{Q} = \text{Proj}_q(\mathbf{X}),\ \mathbf{K} = \text{Proj}_k(\mathbf{X}),\ \mathbf{V} = \text{Proj}_v(\mathbf{X}), \quad (1)$$

$$\mathbf{Y} = \mathbf{X} + \text{Proj}\left(\text{Attn}_{qk} \cdot \mathbf{V}\right), \quad (2)$$

$$\mathbf{Z} = \mathbf{Y} + \text{FFN}(\mathbf{Y}), \quad (3)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent the query, key, and value embeddings, respectively; Proj denotes projection layers; $\text{Attn}_{qk} = \text{SoftMax}\left(\mathbf{Q}\mathbf{K}^\top/\sqrt{d}\right)$ represents the self-attention process, with d denoting the dimension of each attention head. FFN denotes a feed-forward network. For simplicity, normalization operations are omitted.

After passing through the final attention block, $\mathbf{Z}[0]$ represents the global [CLS] token. The remaining image patch embeddings $\mathbf{Z}[1:h\times w]$ can be reshaped to obtain dense feature representations $\mathbf{X}_{\text{dense}} \in \mathbb{R}^{C\times H\times W}$[1].

Adapting CLIP to dense prediction tasks. Several studies have attempted to alleviate the domain shift issue in applying CLIP to dense prediction tasks via fine-tuning strategies. These approaches fall into two main categories:

- **Joint fine-tuning.** These methods fine-tune CLIP while training task-specific components [10, 17, 18, 25, 28, 52, 57]. For instance, CAT-Seg [10] proposes an attention fine-tuning strategy based on ViT CLIP, which general-

[1]The final V-L projection layer is omitted here for brevity.

izes well to unseen categories. MAFT [17] leverages attention bias to fine-tune CLIP for mask classification.

- **Pre-fine-tuning.** These methods directly fine-tune CLIP using cost-efficient techniques [33, 48–50, 63], which are more closely aligned with the approach proposed in this paper. As illustrated in Figure 4(a), CLIM [49] employs a mosaic augmentation technique to stitch multiple images into a single image, enabling each sub-image to serve as a pseudo-region for region-text contrastive learning. CLIP-Self [48] enhances CLIP's region classification accuracy by maximizing cosine similarity between its region representations and the corresponding image crop representations, as illustrated in Figure 4(b).

2.2. Key Observations

Despite the promising results of the two categories of fine-tuned methods in Section 2.1, they continue to exhibit certain limitations. Joint fine-tuning methods are typically specific to tasks or models and heavily rely on labor-intensive annotations of dense prediction tasks. On the other hand, pre-fine-tuning methods demonstrate broader applicability. However, their region-level fine-tuning technique remains limited in image segmentation tasks that require pixel-level details. To tackle this issue, we investigate the feasibility of incorporating pixel-level details into CLIP's pre-fine-tuning, enabling it to better align with open-vocabulary dense prediction tasks. In the following, we start by analyzing CLIP's attention maps across various layers.

The "proxy" token phenomenon. As shown in Figure 3(a), we found that in CLIP's shallow layer, the attention weights of CLIP's [CLS] token are widely distributed across the image (i.e., layer 6). However, in the deeper layers, the [CLS] token shifts its focus away from primary objects in the image and attends to specific tokens, as highlighted by the bright spots within the image background. Additionally, we found that image tokens (rows 2 and 3) exhibit similar behavior to the [CLS] token, showing high attention to certain tokens in the background, regardless of their position.

These background tokens may serve as "proxies" for the [CLS] token. This suggests that these tokens aggregate essential information from other image tokens, enabling the [CLS] token to form an approximate "global view" by summarizing content from them, thereby facilitating image classification. However, these "proxy" tokens negatively affect the feature correlations between image tokens. As illustrated in Figure 3(a), when we shift the position of the anchor image token (from the bird to the branch), we observe that the new image token still pays high attention to the "proxy" tokens. This results in a lack of correlation between image patches that share the same semantics, which is detrimental to dense prediction tasks.

VFMs exhibit better dense correlations. Considering the inherent constraints that impede CLIP's efficacy in dense perception tasks, we instead observe that VFMs such as the DINO series [2, 35], trained in a self-supervised manner, and the SAM series [22, 37], trained on large-scale segmentation data, are capable of extracting features with strong spatial consistency, as shown in Figure 3(b).

Table 1. Performance of different distillation schemes.

Distillation Type	Region Classification (mAcc)		Semantic Segmentation (mIoU)	
	COCO (Thing)	COCO (Stuff)	Context59	CityScape
Self Distillation [48]	69.5	44.6	29.4	25.6
Self+VFM Distillation [22]	65.6 (-3.9)	41.3 (-3.3)	32.4 (+3.0)	28.7 (+3.1)
Self+VFM+Decouple	75.0 (+5.5)	51.8 (+7.2)	35.3 (+5.9)	32.3 (+6.7)

In particular, the attention map of VFMs does not exhibit the "proxy" token phenomenon observed in CLIP. Furthermore, when we change the position of the anchor image token, the VFM shows a better correlation for image tokens with the same semantics. Therefore, we consider whether VFMs can be incorporated into the pre-fine-tuning process to further improve the feature correlations of CLIP. However, this straightforward approach fails to achieve satisfactory results. Conducting VFM distillation[2] and self-distillation[3] simultaneously results in reduced region classification performance, as shown in Table 1 (row 2). We hypothesize that this observation stems from the fact that spatial feature correlation and vision-language alignment have different optimization focuses, and optimizing them simultaneously within a single model results in trade-offs.

3. Method

Through the above analysis, we found that CLIP underperforms in dense prediction tasks since its image tokens fail to effectively aggregate information from semantically related regions. Observations of VFMs' attention maps inspired us to incorporate them into CLIP's pre-fine-tuning process. Considering the optimization conflict between feature correlations and visual-language alignment, we applied a decoupled feature enhancement strategy to CLIP.

In this section, we introduce DeCLIP, an unsupervised fine-tuning framework for adapting CLIP to dense prediction tasks. We first explain how to decouple CLIP's self-attention mechanism into "content" and "context" components in Sec.3.1, then describe how these components learn from different "teacher" models in Sec.3.2 by distillation.

3.1. Decoupled Attention

The unsuccessful attempts to simultaneously perform self-distillation and VFM distillation on $\mathbf{X}_{\text{dense}}$ (Table 1, row 2) prompted us to explore the feasibility of a decoupled distillation. In the following, we propose decoupling CLIP's

[2] "VFM distillation" indicates aligning the feature self-correlations between CLIP's $\mathbf{X}_{\text{dense}}$ and that of the VFM.

[3] "Self-distillation" refers to aligning region features from $\mathbf{X}_{\text{dense}}$ with their corresponding [CLS] representation.

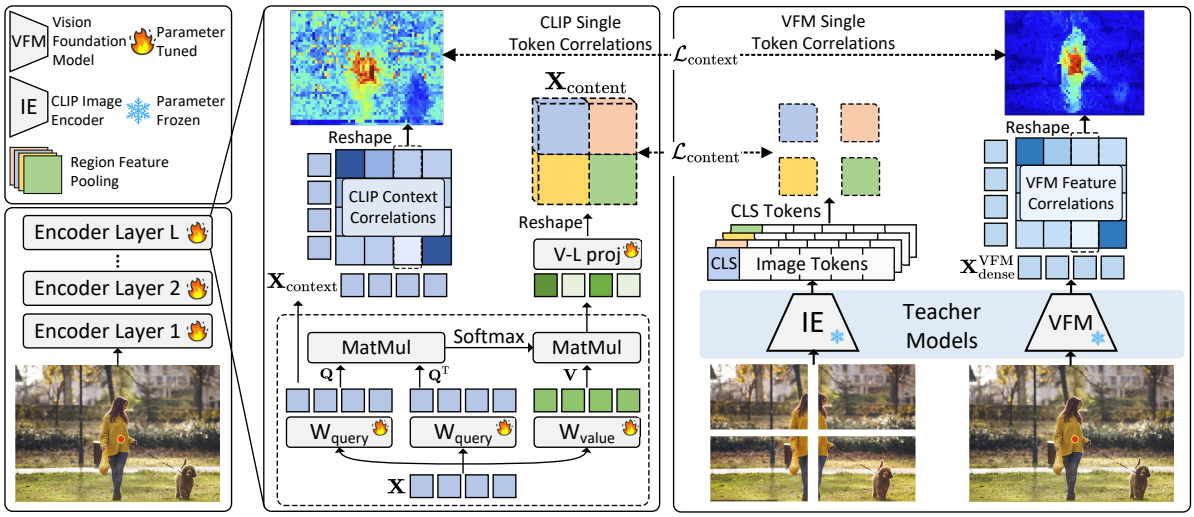

Figure 5. **Illustration of the DeCLIP framework.** We decouple CLIP's final attention module into context and content features for distillation, avoiding optimization conflicts between feature correlations and visual-language alignment. CLIP itself serves as the teacher for content features to improve region classification accuracy. A VFM serves as the teacher for context features to enhance spatial consistency.

self-attention module to obtain "content" and "context" features, and separately optimize the local discriminability and spatial consistency abilities, as illustrated in Figure 4(c).

Rethinking the self-attention. As described in Sec.2.1, in CLIP's last attention block, the $\mathbf{V}$ features are weighted and summed under the guidance of the attention map (Attn_{qk}) derived from $\mathbf{Q}$ and $\mathbf{K}$, which define spatial or semantic relationships among image tokens. Studies [24, 41, 45, 51] have shown that CLIP's dense features $\mathbf{X}_{\text{dense}}$ can be directly used for semantic segmentation by per-pixel classification, indicating that each pixel of $\mathbf{X}_{\text{dense}}$ contains independent semantic information. Inspired by this, we regard $\mathbf{Q}$ and $\mathbf{K}$ as anchors for improving spatial consistency, and $\mathbf{X}_{\text{dense}}$ as an anchor for enhancing local discriminability.

Additionally, recent training-free OVS studies [24, 45] have further promoted us to decouple CLIP's self-attention followed by distillation. They modify CLIP's attention block from Attn_{qk} to Attn_{qq} and remove the residual connections, simplifying the optimization of local feature consistency by focusing on $\mathbf{Q}$ alone. Based on our rethinking of CLIP's self-attention and inspired by these methods, we propose decoupling CLIP's last attention block to obtain "content" and "context" features for distillation as follows:

$$\mathbf{X}_{\text{context}} = \text{Proj}_q(\mathbf{X}),\ \mathbf{V} = \text{Proj}_v(\mathbf{X}), \tag{4}$$

$$\mathbf{X}_{\text{content}} = \text{Proj}\left(\text{Attn}_{\text{context}} \cdot \mathbf{V}\right), \tag{5}$$

$$\text{Attn}_{\text{context}} = \text{SoftMax}\left(\mathbf{X}_{\text{context}}\mathbf{X}_{\text{context}}^{\top}/\sqrt{d}\right). \tag{6}$$

Specifically, $\mathbf{V}$ is aggregated based on the attention map ($\text{Attn}_{\text{context}}$) generated from $\mathbf{X}_{\text{context}}$. $\mathbf{X}_{\text{context}}$ determines which image tokens are semantically or spatially related. $\mathbf{X}_{\text{content}}$ carries the semantic information of each image token in the visual-language space. By decoupling the features in this manner, we can apply different guidance constraints to $\mathbf{X}_{\text{context}}$ and $\mathbf{X}_{\text{content}}$ to obtain diverse feature representations in a unified architecture without interference.

As observed in Sec. 2.2, VFM exhibits a strong correlation for image tokens with the same semantics, thus we leverage it as guidance for $\mathbf{X}_{\text{context}}$ to improve CLIP's local feature spatial consistency. Meanwhile, we employ the self-distillation technique as guidance for $\mathbf{X}_{\text{content}}$ to enhance the visual-language alignment of CLIP's region feature.

As demonstrated in Table 1 (row 3), this decoupled optimization significantly improves the local discriminability and spatial consistency of CLIP's features, leading to simultaneous enhancements in both region classification accuracy and semantic segmentation performance.

3.2. DeCLIP

The previous section presents a method for obtaining the decoupled "context" and "content" features from CLIP. In this section, we elaborate on how the decoupled features $\mathbf{X}_{\text{content}}$ and $\mathbf{X}_{\text{context}}$ learn from their respective teacher models to enhance CLIP's performance on open-vocabulary dense prediction tasks.

Content feature distillation. As shown in Figure 5, the first teacher model in DeCLIP is itself, which is known as self-distillation [6, 33, 34, 48]. we employ an image patching method to align the region representations of the student model's feature map with the corresponding image crop representations (i.e., [CLS] token) of the teacher model.

Specifically, the input image $\mathbf{I}$ is first divided into k sub-

regions. Subsequently, these sub-regions are cropped from the original image, resulting in a set of sub-images $S = \{\mathbf{I}'_1, \mathbf{I}'_2, \ldots, \mathbf{I}'_k\}$. The student model takes the image $\mathbf{I}$ as input and outputs the content feature $\mathbf{X}_{\text{content}} \in \mathbb{R}^{C\times H\times W}$ and the context feature $\mathbf{X}_{\text{context}} \in \mathbb{R}^{D\times H\times W}$, as mentioned in Eq.(6). Here, D represents the dimension of the CLIP visual encoder, and C represents the shared dimension of the vision-language modality. Then, the student model uses RoI Align [16] to pool region features from $\mathbf{X}_{\text{content}}$ based on the cropping coordinates of S, resulting in a region feature set $F_s = \{\boldsymbol{f}_1^s, \boldsymbol{f}_2^s, \ldots, \boldsymbol{f}_k^s\}$, where $\boldsymbol{f}_i^s \in \mathbb{R}^{1\times C}$.

Meanwhile, the teacher model takes the sub-image set S as input and outputs a series of [CLS] tokens corresponding to the cropped sub-images, resulting in [CLS] token set $F_t = \{\boldsymbol{f}_1^t, \boldsymbol{f}_2^t, \ldots, \boldsymbol{f}_k^t\}$, where $\boldsymbol{f}_i^t \in \mathbb{R}^{1\times C}$. Finally, we use a cosine similarity loss to align the [CLS] tokens from F_t with the region features from F_s as follows:

$$\mathcal{L}_{\text{content}} = \frac{1}{k}\sum_{i=1}^{k}\left(1 - \frac{\boldsymbol{f}_i^t \cdot \boldsymbol{f}_i^s}{\|\boldsymbol{f}_i^t\| \cdot \|\boldsymbol{f}_i^s\|}\right). \tag{7}$$

The intuition behind this distillation branch is that, for objects within an image, classifying them using image crops (i.e., [CLS] token) achieves higher accuracy than using region features [48]. This is because CLIP is pre-trained on image-text pairs using contrastive learning, as mentioned in Sec.2.1. Therefore, the distillation learning of $\mathbf{X}_{\text{content}}$ enhances the discriminability of CLIP's region features, *i.e.*, $F_s = \{\boldsymbol{f}_1^s, \boldsymbol{f}_2^s, \ldots, \boldsymbol{f}_k^s\}$, by mimicking the [CLS] tokens obtained from the image crops, *i.e.*, $F_t = \{\boldsymbol{f}_1^t, \boldsymbol{f}_2^t, \ldots, \boldsymbol{f}_k^t\}$. However, as previously discussed in Sec.2.2, the region-level fine-tuning remains limited in image segmentation that requires pixel-wise scene understanding.

Context feature distillation. As discussed in Sec.2.2, VFMs do not exhibit CLIP's "proxy" token issue and better correlate semantically related image tokens, which may be conducive to the fine-grained local perception. Therefore, we distilled these correlations into CLIP's $\mathbf{X}_{\text{context}}$ features.

As illustrated in Figure 5, the same image $\mathbf{I}$ is input into the VFM to obtain its dense feature representations $\mathbf{X}_{\text{dense}}^{\text{VFM}} \in \mathbb{R}^{D\times HW}$. To ensure consistency in the number of image tokens after patch embedding, different input resolutions are typically used for the VFM and the student CLIP. To transfer VFM's correlations between image tokens to CLIP, an intermediary is required to represent the correlation volume between two image tokens. Cosine similarity is used in our method, specifically as follows:

$$r_{ij} = \frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j}{\|\boldsymbol{x}_i\| \cdot \|\boldsymbol{x}_j\|}. \tag{8}$$

Here, $\boldsymbol{x}_i \in \mathbb{R}^{1\times D}$ and $\boldsymbol{x}_j \in \mathbb{R}^{1\times D}$ represent the i-th and j-th image patch tokens. r_{ij} denotes the correlation volume between patch tokens $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We use the L2 loss to align the discrepancy in the correlation volume between the image tokens of $\mathbf{X}_{\text{dense}}^{\text{VFM}}$ and $\mathbf{X}_{\text{context}}$, specifically as follows:

$$\mathcal{L}_{\text{context}} = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\left\|r_{ij}^{\text{VFM}} - r_{ij}^{\text{CLIP}}\right\|_2, \tag{9}$$

where r_{ij}^{VFM} and r_{ij}^{CLIP} denote the correlation volume between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for VFM and CLIP, respectively. Finally, the entire distillation learning process of DeCLIP can be expressed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{content}} + \lambda\mathcal{L}_{\text{context}}, \tag{10}$$

where λ[4] represents the loss scaling hyperparameter.

4. Experiments

4.1. Datasets and Evaluation

We conducted extensive evaluations across multiple open-vocabulary dense prediction benchmarks, encompassing object detection, semantic segmentation, and segmentation based on VLM features. Due to space limitations, detailed descriptions of the datasets, evaluation metrics, and implementation specifics are provided in the Appendix.

4.2. Benchmark Results

Open-Vocabulary Detection. Table 2 presents DeCLIP's performance on OV-COCO and OV-LVIS benchmarks. On OV-COCO, DeCLIP improves the F-ViT [48] baseline by 3.5 and 1.9 mAP, and the OV-DQUO [46] baseline by 6.9 and 2.7 mAP on novel classes. On OV-LVIS, it achieves gains of 1.5 and 2.3 mAP with F-ViT, as well as 1.3 and 2.2 mAP with OV-DQUO on rare classes. Cross-dataset evaluations of F-ViT+DeCLIP trained on OV-LVIS (Table 3) further confirm DeCLIP's superiority over existing methods.

Open-Vocabulary Semantic Segmentation. Table 4 displays the performance of the CAT-Seg [10] model using DeCLIP as the backbone across various open-vocabulary semantic segmentation benchmarks. The results show that DeCLIP significantly enhances segmentation performance on all datasets. Notably, even with the ViT-B/16 version of DeCLIP, CAT-Seg nearly surpasses all existing SOTA methods that utilize substantially larger encoders like ConvNeXt-L. When employing the ViT-L/14 version of DeCLIP, the model achieves new SOTA results in open-vocabulary semantic segmentation tasks.

Open-Vocabulary Semantic Segmentation Based on VLM Features. Following existing methods [24, 41, 45], in this experiment, we assign each pixel in the feature map the category with which it has the highest cosine similarity. The low-resolution prediction result is up-sampled to the

[4]The sensitivity analysis is in the appendix.

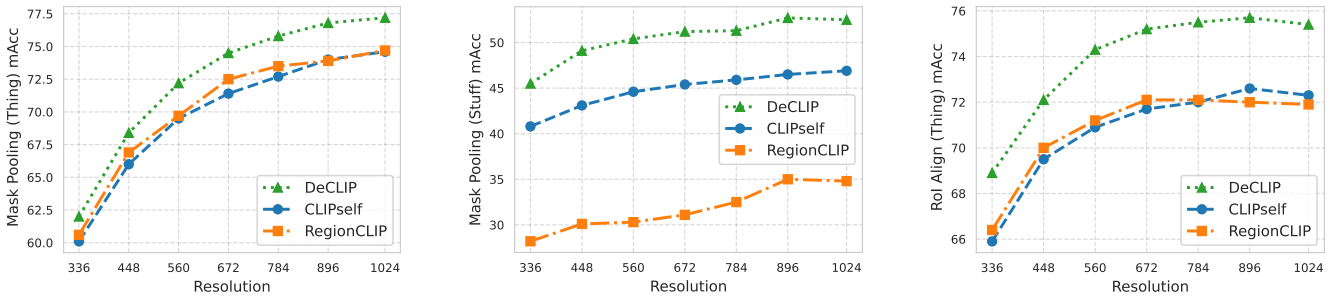

Figure 6. Comparisons between DeCLIP and existing methods in terms of open-vocabulary region classification ability at different resolutions on the COCO panoptic dataset.

Table 2. Comparison with state-of-the-art open-vocabulary object detection methods. Caption supervision indicates that the method learns from extra image-text pairs, while CLIP supervision refers to transferring knowledge from CLIP. †: DETR-based detectors [1].

(a) OV-COCO benchmark

Method	Supervision	Backbone	AP_{50}^{Novel}
ViLD [14]	CLIP	RN50	27.6
Detic [65]	Caption	RN50	27.8
OV-DETR† [60]	CLIP	RN50	29.4
BARON-KD [47]	CLIP	RN50	34.0
SAS-Det [62]	CLIP	RN50	37.4
OV-DQUO† [46]	CLIP	RN50	39.2
RegionCLIP [63]	Captions	RN50x4	39.3
CORA† [50]	CLIP	RN50x4	41.7
OV-DQUO† [46]	CLIP	RN50x4	45.6
RO-ViT [20]	CLIP	ViT-L/16	33.0
CFM-ViT [19]	CLIP	ViT-L/16	34.1
CLIPSelf [48]	CLIP	ViT-B/16	37.6
CLIPSelf [48]	CLIP	ViT-L/14	44.3
F-ViT [48]+DeCLIP	CLIP	ViT-B/16	41.1 (+3.5)
F-ViT [48]+DeCLIP	CLIP	ViT-L/14	46.2 (+1.9)
OV-DQUO+DeCLIP†	CLIP	ViT-B/16	**46.1** (+6.9)
OV-DQUO+DeCLIP†	CLIP	ViT-L/14	**48.3** (+2.7)

(b) OV-LVIS benchmark

Method	Supervision	Backbone	mAP_r
ViLD [14]	CLIP	RN50	16.3
OV-DETR† [60]	CLIP	RN50	17.4
BARON-KD [47]	CLIP	RN50	22.6
RegionCLIP [63]	Caption	RN50x4	22.0
OV-SAM [59]	CLIP	RN50x16	24.0
CORA+† [50]	Caption	RN50x4	28.1
F-VLM [23]	CLIP	RN50x64	32.8
CLIPSelf [48]	CLIP	ViT-B/16	25.3
OV-DQUO† [46]	CLIP	ViT-B/16	29.7
Detic [65]	Caption	Swin-B	33.8
RO-ViT [20]	CLIP	ViT-H/16	34.1
CLIPSelf [48]	CLIP	ViT-L/14	34.9
OV-DQUO† [46]	CLIP	ViT-L/14	39.3
F-ViT [48]+DeCLIP	CLIP	ViT-B/16	26.8 (+1.5)
F-ViT [48]+DeCLIP	CLIP	ViT-L/14	37.2 (+2.3)
OV-DQUO+DeCLIP†	CLIP	ViT-B/16	**31.0** (+1.3)
OV-DQUO+DeCLIP†	CLIP	ViT-L/14	**41.5** (+2.2)

Table 3. Transfer evaluation of the LVIS-trained detector on COCO and Objects365 datasets.

Method	COCO			Objects365 [40]		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Supervised Baseline [14]	46.5	67.6	50.9	25.6	38.6	28.0
ViLD [14]	36.6	55.6	39.6	11.8	18.0	12.6
DetPro [13]	34.9	53.8	37.4	12.1	18.8	12.9
BARON [47]	36.2	55.7	39.1	13.6	21.0	14.5
F-VLM [23]	37.9	61.6	41.2	16.2	27.4	17.5
CoDet [32]	39.1	57.0	42.3	14.2	20.5	15.3
RO-ViT [21]	-	-	-	17.7	27.4	19.1
CLIPSelf [48]	40.5	63.8	44.3	19.5	31.3	20.7
DeCLIP	**41.0**	**64.6**	**44.8**	**20.0**	**32.2**	**21.2**

original resolution to obtain the final segmentation map. As shown in Table 5, DeCLIP outperforms all existing methods in terms of average mIoU across eight benchmarks, highlighting the effectiveness of our approach in improving the discriminability and spatial consistency of VLM features.

Open-Vocabulary Region Classification. We assess the region classification performance of DeCLIP, RegionCLIP [63], and CLIPSelf [48] at various resolutions on the COCO-Panoptic validation set. Using RoI Align [16] and Mask Pooling, we extract local features from the feature maps based on annotated bounding boxes and masks, assigning categories based on maximum cosine similarity. As illustrated in Figure 6, the Top-1 mean accuracy (mAcc) results demonstrate that DeCLIP consistently surpasses existing methods in region recognition across all resolutions.

4.3. Ablation Study

The impact of VFMs. We analyzed the impact of various VFM configurations on DeCLIP performance. As shown in Table 6, DeCLIP distilled from DINO [2] performs moderately in segmentation but trails SAM [22, 37] and DINOv2 [35] in region classification. DeCLIP distilled from SAM excels in region classification but shows lower segmentation performance compared to DINO. DINOv2 achieves balance in both region classification and segmentation.

Table 4. Results on open-vocabulary semantic segmentation. † indicates results re-experimented by CAT-Seg [10].

Method	Backbone	Training Set	ADE847	Context459	ADE150	Context59	VOC20	VOC21
ZegFormer† [11]	ViT-B/16	COCO-Stuff	5.6	10.4	18.0	45.5	89.5	65.5
ZSseg [55]	ViT-B/16	COCO-Stuff	7.0	-	20.5	47.7	88.4	-
OVSeg [28]	ViT-L/14	COCO-Stuff	9.0	12.4	29.6	55.7	94.5	-
SAN [56]	ViT-L/14	COCO-Stuff	13.7	17.1	33.3	60.2	95.5	-
ODISE [54]	ViT-L/14	COCO-Panoptic	11.1	14.5	29.9	57.3	-	84.6
MAFT [17]	ConvNeXt-L	COCO-Stuff	13.1	17.0	34.4	57.5	93.0	-
FC-CLIP [58]	ConvNeXt-L	COCO-Panoptic	14.8	18.2	34.1	58.4	95.4	81.8
FrozenSeg [7]	ConvNeXt-L	COCO-Panoptic	14.8	19.7	34.4	-	-	82.5
CAT-Seg [10]	ViT-B/16	COCO-Stuff	12.0	19.0	31.8	57.5	94.6	77.3
CAT-Seg [10]	ViT-L/14	COCO-Stuff	16.0	23.8	37.9	63.3	97.0	82.5
CAT-Seg+DeCLIP	ViT-B/16	COCO-Stuff	15.3 (+3.3)	21.4 (+2.4)	36.3 (+4.5)	60.6 (+3.1)	96.6 (+2.0)	81.3 (+4.0)
CAT-Seg+DeCLIP	ViT-L/14	COCO-Stuff	**17.6** (+1.6)	**25.9** (+2.1)	**40.7** (+2.8)	**63.9** (+0.6)	**97.7** (+0.7)	**83.9** (+1.4)

Table 5. Results on open-vocabulary semantic segmentation based on VLM features.

Method	With a background category			Without background category					Avg
	VOC21	Context60	COCO-Obj	VOC20	CityScape	Context59	ADE	COCO-Stf	
CLIP [36]	18.8	9.9	8.1	49.4	6.5	11.1	3.1	5.7	14.1
MaskCLIP [64]	43.4	23.2	20.6	74.9	24.9	26.4	11.9	16.7	30.3
GroupViT [53]	52.3	18.7	27.5	79.7	18.5	23.4	10.4	15.3	30.7
ReCo [42]	25.1	19.9	15.7	57.7	21.6	22.3	11.2	14.8	23.5
TCL [3]	51.2	24.3	30.4	77.5	23.5	30.3	14.9	19.6	33.9
OVSeg [28]	53.8	20.4	25.1	-	-	-	5.6	-	-
SCLIP [45]	59.1	30.4	30.5	80.4	32.2	34.2	16.1	22.4	38.2
ClearCLIP [24]	51.8	32.6	33.0	80.9	30.0	35.9	16.7	23.9	38.1
CLIP-DINOiser [51]	**62.1**	32.4	34.8	80.9	31.7	35.9	20.0	24.6	40.3
DeCLIP (ours)	59.7	**35.3**	**36.4**	**85.0**	**32.8**	**39.2**	**21.9**	**25.3**	**41.9**

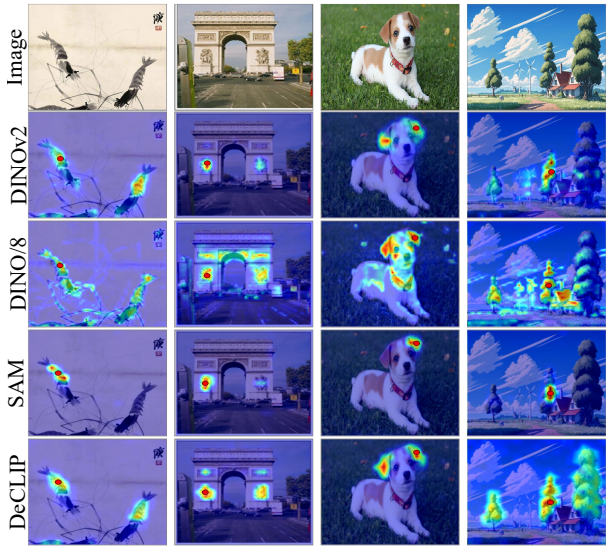

Figure 7. Qualitative comparisons of attention maps between VFMs and DeCLIP. The anchor image token is marked in red.

Qualitative results. Figure 7 presents the visual comparison of attention maps between DINO, SAM, DINOv2, and DeCLIP. Experimental results show that DeCLIP effectively focuses on regions spatially or semantically associated with the anchor image token. Moreover, this experiment reveals why DeCLIP distilled from DINOv2 works best: SAM lacks semantic association ability, while DINO focus indiscriminately on all primary objects in the image.

Table 6. Ablation studies on the impact of different VFMs on open-vocabulary region classification and segmentation.

VFMs	Arch	Region Classification (mAcc)		Semantic Segmentation (mIoU)		
		COCO (Thing)	COCO (Stuff)	Context59	COCO-Stf	ADE
DINO [2]	ViT-B/8	68.4	49.4	37.3	23.2	19.5
DINO [2]	ViT-B/16	67.6	47.4	38.1	23.7	20.4
SAM [22]	ViT-B/16	75.0	51.8	35.3	22.0	18.5
SAM [22]	ViT-L/16	76.8	52.6	37.7	23.0	20.0
DINOv2 [35]	ViT-B/14	77.2	52.5	**39.2**	**25.3**	**21.9**
DINOv2 [35]	ViT-L/14	**77.6**	**53.1**	38.0	24.1	21.3

5. Conclusion

This paper analyzes the limitations of CLIP in dense prediction tasks from the perspective of its attention map. We observed that CLIP's [CLS] token negatively affects the attention map of image tokens. To address this issue, we proposed DeCLIP, a decoupled feature enhancement strategy. Extensive experiment results on open-vocabulary dense prediction benchmarks demonstrate that DeCLIP outperforms state-of-the-art methods, achieving excellent performance across all evaluated task domains.

Acknowledgement. This work was supported by Guangdong Basic and Applied Basic Research Foundation (2025A1515011546), by Shenzhen Science and Technology Innovation Program (JCYJ20240813105901003, KJZD20240903102901003), and by Science and Technology Project of Shenzhen (GXWD-20220811170603002)

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020. 7

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660, 2021. 4, 7, 8

[3] Junbum Cha, Jonghwan Mun, and Byungseok Roh. 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, pages 11165–11174, 2023. 8

[4] Fangyi Chen, Han Zhang, Kai Hu, Yu-Kai Huang, Chenchen Zhu, and Marios Savvides. Enhanced training of query-based object detection via selective query recollection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 23756–23765, 2023. 1

[5] Fangyi Chen, Han Zhang, Zhantao Yang, Hao Chen, Kai Hu, and Marios Savvides. Rtgen: Generating region-text pairs for open-vocabulary object detection. arXiv preprint arXiv:2405.19854, 2024. 1

[6] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 699–710, 2023. 5

[7] Xi Chen, Haosen Yang, Sheng Jin, Xiatian Zhu, and Hongxun Yao. Frozenseg: Harmonizing frozen foundation models for open-vocabulary segmentation. arXiv preprint arXiv:2409.03525, 2024. 1, 8

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 1290–1299, 2022. 1

[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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2818–2829, 2023. 1

[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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4113–4123, 2024. 1, 3, 6, 8

[11] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11583–11592, 2022. 8

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020. 2

[13] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14084–14093, 2022. 7

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921, 2021. 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 6, 7

[17] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. Advances in Neural Information Processing Systems, 36:35631–35653, 2023. 1, 3, 4, 8

[18] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In European Conference on Computer Vision, pages 399–416. Springer, 2025. 3

[19] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Contrastive feature masking open-vocabulary vision transformer. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pages 15556–15566, 2023. 7

[20] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11144–11154, 2023. 7

[21] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 11144–11154, 2023. 7

[22] 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 Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4015–4026, 2023. 4, 7, 8

[23] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. arXiv preprint arXiv:2209.15639, 2022. 1, 7

[24] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. arXiv preprint arXiv:2407.12442, 2024. 5, 6, 8

[25] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546, 2022. 3

[26] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3041–3050, 2023. 1

[27] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23390–23400, 2023. 1

[28] 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 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7061–7070, 2023. 3, 8

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 2

[30] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In International Conference on Learning Representations, 2022. 1

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 2

[32] Chuofan Ma, Yi Jiang, Xin Wen, Zehuan Yuan, and Xiaojuan Qi. Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. Advances in Neural Information Processing Systems, 36, 2024. 7

[33] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19413–19423, 2023. 4, 5

[34] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In European Conference on Computer Vision, pages 38–55. Springer, 2025. 5

[35] 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. 4, 7, 8

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, pages 8748–8763. PMLR, 2021. 1, 2, 8

[37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. 4, 7

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28, 2015. 1

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, pages 234–241. Springer, 2015. 1

[40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In Proceedings of the IEEE/CVF international conference on computer vision, pages 8430–8439, 2019. 7

[41] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In European Conference on Computer Vision, pages 139–156. Springer, 2025. 5, 6

[42] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. Advances in Neural Information Processing Systems, 35:33754–33767, 2022. 8

[43] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023. 1

[44] A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017. 2

[45] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. arXiv preprint arXiv:2312.01597, 2023. 5, 6, 8

[46] Junjie Wang, Bin Chen, Bin Kang, Yulin Li, YiChi Chen, Weizhi Xian, and Huifeng Chang. Ov-dquo: Open-vocabulary detr with denoising text query training and open-world unknown objects supervision. arXiv preprint arXiv:2405.17913, 2024. 1, 6, 7

[47] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15254–15264, 2023. 1, 7

[48] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In The Twelfth International Conference on Learning Representations, 2024. 1, 4, 5, 6, 7

[49] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Wentao Liu, and Chen Change Loy. Clim: Contrastive language-image mosaic for region representation. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 6117–6125, 2024. 1, 4

[50] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7031–7040, 2023. 1, 4, 7

[51] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks. arXiv preprint arXiv:2312.12359, 2023. 5, 8

[52] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3426–3436, 2024. 3

[53] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18134–18144, 2022. 8

[54] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2955–2966, 2023. 8

[55] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In European Conference on Computer Vision, pages 736–753. Springer, 2022. 8

[56] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2945–2954, 2023. 8

[57] Xin Xu, Tianyi Xiong, Zheng Ding, and Zhuowen Tu. Masq-clip for open-vocabulary universal image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 887–898, 2023. 3

[58] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. Advances in Neural Information Processing Systems, 36, 2024. 1, 8

[59] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. In ECCV, 2024. 7

[60] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In European Conference on Computer Vision, pages 106–122. Springer, 2022. 7

[61] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14393–14402, 2021. 1

[62] Shiyu Zhao, Samuel Schulter, Long Zhao, Zhixing Zhang, Vijay Kumar B G, Yumin Suh, Manmohan Chandraker, and Dimitris N. Metaxas. Taming self-training for open-vocabulary object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13938–13947, 2024. 1, 7

[63] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16793–16803, 2022. 4, 7

[64] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In European Conference on Computer Vision, pages 696–712. Springer, 2022. 8

[65] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In European Conference on Computer Vision, pages 350–368. Springer, 2022. 7

[66] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020. 1