Mask-Adapter: The Devil is in the Masks for Open-Vocabulary Segmentation

Yongkang Li*, Tianheng Cheng*, Bin Feng, Wenyu Liu, Xinggang Wang†
School of EIC, Huazhong University of Science & Technology
{liyk, thch, fengbin, liuwy, xgwang}@hust.edu.cn

Abstract

Recent open-vocabulary segmentation methods adopt mask generators to predict segmentation masks and leverage pre-trained vision-language models, e.g.*, CLIP, to classify these masks via mask pooling. Although these approaches show promising results, it is counterintuitive that accurate masks often fail to yield accurate classification results through pooling CLIP image embeddings within the mask regions. In this paper, we reveal the performance limitations of mask pooling and introduce **Mask-Adapter**, a simple yet effective method to address these challenges in open-vocabulary segmentation. Compared to directly using proposal masks, our proposed Mask-Adapter extracts semantic activation maps from proposal masks, providing richer contextual information and ensuring alignment between masks and CLIP. Additionally, we propose a mask consistency loss that encourages proposal masks with similar IoUs to obtain similar CLIP embeddings to enhance models' robustness to varying predicted masks. Mask-Adapter integrates seamlessly into open-vocabulary segmentation methods based on mask pooling in a plug-and-play manner, delivering more accurate classification results. Extensive experiments across several zero-shot benchmarks demonstrate significant performance gains for the proposed Mask-Adapter on several well-established methods. Notably, Mask-Adapter also extends effectively to SAM and achieves impressive results on several open-vocabulary segmentation datasets. Code and models are available at* `https://github.com/hustvl/MaskAdapter`.

1. Introduction

Image segmentation, which aims to assign each pixel a label from a predefined set of categories, has seen significant advancements. Despite the success of numerous image segmentation approaches [5–7, 35, 38, 42, 51], these approaches cannot segment categories unseen during training, limiting their applicability in open-world scenarios. Open-vocabulary segmentation (OVS) has expanded traditional image segmentation from fixed-set to open-set, enabling the segmentation of categories not appeared in the training set. In contrast to fixed-category segmentation, open-vocabulary segmentation not only supports open-set segmentation but also enables challenging segmentation based on arbitrary text inputs, *e.g.*, categories and descriptions.

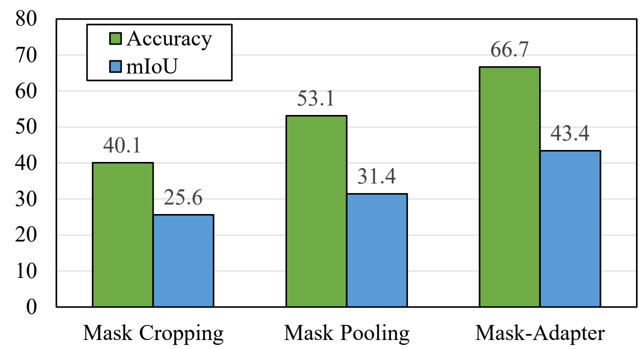
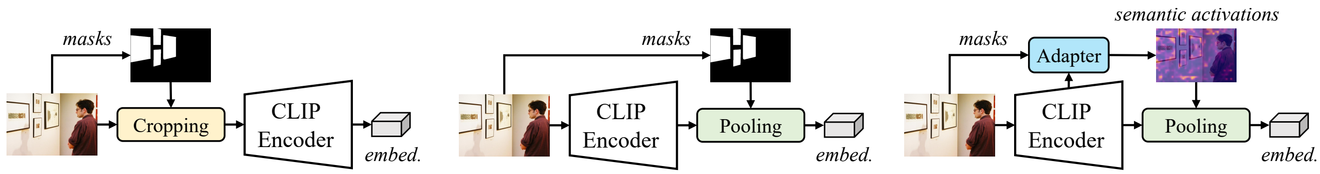


Figure 1. **Upper bound analysis of mask embedding extraction methods.** Using ADE20K [52] ground-truth masks as input, we evaluate the upper bound of different extraction methods. Although mask cropping and mask pooling show limited performance with ground-truth masks, our Mask-Adapter significantly enhances the upper bound for open-vocabulary segmentation.

Recent research tackles open-vocabulary segmentation by leveraging pre-trained vision-language models, *e.g.*, CLIP [30] and ALIGN [17], which are pre-trained on large-scale image-text pairs and have strong zero-shot recognition capabilities. Notably, several works [10, 16, 19, 23, 32, 44, 46, 48], such as SimpleBaseline [44] and FC-CLIP [48], follow a *segment-then-recognize* paradigm for open-vocabulary segmentation, which adopts mask generators to predict proposal masks from images and then classify the proposal masks by transferring the zero-shot recognition capability from the frozen CLIP. Recognizing segmentation masks is primarily achieved by extracting mask embeddings through the CLIP visual encoder and matching them with text embeddings. Specifically, the mask embeddings can be obtained in two ways: (1) *mask cropping* [10, 23, 44]: cropping the segmented region from

* Equal contribution.
† Corresponding author: Xinggang Wang



(a) Mask Cropping (b) Mask Pooling (c) Mask-Adapter

Figure 2. **Comparison of mask embedding extraction methods.** (a) **Mask Cropping**: cropping the segmented region from the image and feeding it into CLIP to extract mask embeddings. (b) **Mask Pooling**: aggregating region features with the proposal masks. (c) **Mask-Adapter**: the proposal masks and CLIP features are passed through Mask-Adapter to extract semantic activation maps, which are then used to construct mask embeddings by aggregating CLIP features based on these highlighted regions and contextual information.

the image and using CLIP to extract image embeddings; (2) *mask pooling* [15, 19, 32, 46, 48]: directly aggregating the features of the region using mask pooling [15] with the proposal masks through dot-product, which is more efficient than mask cropping and can be end-to-end optimized. However, the upper bound of these two methods for open-vocabulary segmentation is inherently limited, as shown in Fig. 1. The mask cropping method (Fig. 2 (a)) fails to consider the substantial differences between the masked image and the natural images used during CLIP [30] pre-training, as mentioned in [23]. Meanwhile, the mask pooling method (Fig. 2 (b)) fails to capture semantic details and contextual information during feature aggregation, limiting its upper bound.

In this paper, we propose a simple yet effective method to address the bottlenecks in open-vocabulary segmentation. Instead of directly using predicted masks and mask pooling (or mask cropping) to extract mask embeddings, we introduce a *Mask-Adapter* (Fig. 2 (c)) to extract *semantic activation maps* from proposal masks and CLIP features, inspired by SparseInst [7]. Mask embeddings are aggregated based on both semantic and contextual information and subsequently matched with text embeddings for mask recognition. Fig. 2 compares mask cropping, mask pooling, and Mask-Adapter methods. In contrast to previous approaches, Mask-Adapter offers several key advantages: (1) rather than neglecting the background, it aggregates mask embeddings across the entire image, incorporating contextual information to enrich the feature representation; (2) unlike mask pooling, which only conveys the positional information of target regions, semantic activation maps selectively highlights informative regions relevant to recognition, while suppressing less informative areas, thus enhancing feature discriminability; (3) our method preserves the generalization capabilities of CLIP during training while concurrently improving mask recognition performance. Furthermore, we introduce a mask consistency loss to enforce similar proposal masks to obtain similar CLIP mask embeddings, thereby improving robustness to variations in predicted masks. To further mitigate overfitting, we replace the Hungarian matcher with an IoU-based matcher, enabling the model to train on a broader range of proposal masks. In addition, we propose a stable mask-text alignment training strategy that enhances the robustness of the Mask-Adapter.

The main contributions of this paper can be summarized as follows:

- We introduce the Mask-Adapter, which addresses critical bottlenecks, *i.e.*, misalignment between pre-trained CLIP and segmentation masks, in open-vocabulary segmentation and can be integrated seamlessly into mask pooling-based methods in a plug-and-play manner, delivering more accurate classification results.
- We introduce a mask consistency loss and present an IoU-based matcher to reduce overfitting and improve the robustness of Mask-Adapter along with a stable mask-text alignment training strategy.
- We integrate the proposed Mask-Adapter into established open-vocabulary segmentation frameworks, demonstrating substantial performance improvements and achieving state-of-the-art results on datasets such as ADE20K and Pascal-Context. Furthermore, the Mask-Adapter effectively extends to SAM without training, achieving impressive results across multiple open-vocabulary segmentation benchmarks.

2. Related Work

2.1. Open-Vocabulary Segmentation

Open-vocabulary segmentation aims to segment images based on arbitrary textual descriptions rather than predefined classes, primarily categorizing into two types: per-pixel classification [9, 21, 41] and mask classification [10, 15, 16, 19, 23, 43, 44, 48]. OpenSeg [15] first employs mask pooling method to aggregate region features for mask classification. SimpleBaseline [44] and Zegformer [10] propose two-stage frameworks which generate class-agnostic masks and feed the masked image into CLIP for classification. Per-pixel methods, *e.g.*, CAT-Seg [9] and SED [41], adopt cost aggregation to refine the pixel-text cost volume and predict per-pixel classification outputs. PrimitiveNet [25] decomposes complex language constraints into simple primitives and employs cross-primitive and language-primitive attention to enhance referring segmentation performance. SAN [45] attaches an adapter to CLIP, enabling mask proposal generation and recognition. FC-CLIP [48] utilizes CLIP as a shared backbone for both mask generation and mask classification. However, both mask cropping and mask pooling approaches used in previous mask classification methods to obtain mask embeddings exhibit a limited upper bound performance, as illustrated in Fig. 1.

Several works [8, 23] point out that the bottleneck of open-vocabulary segmentation lies in CLIP-based mask classification. OVSeg [23] observes the significant gap between masked images and natural images when using CLIP and addresses it by adapting CLIP with masked image-category pairs. MaskCLIP [11] and MasQCLIP [47] adopt masked attention instead of mask cropping for mask classification. MAFT [18] fine-tunes the CLIP image encoder while employing a self-distillation loss to prevent catastrophic forgetting, and MAFTP [19] further introduces content-dependent transfer to fine-tune the text representations. Deop [16] decouples mask generation and classification and employs a heatmap-based pooling method instead of mask pooling, but its open-vocabulary segmentation performance remains relatively weak. In contrast, our proposed Mask-Adapter extracts *semantic activation maps* from proposal masks, providing richer contextual information and ensuring better alignment between masks and CLIP. Compared to Deop, our method differs in the following key aspects: (1) our Mask-Adapter is a standalone module that can be integrated seamlessly into open-vocabulary segmentation frameworks and SAM; (2) the structure of our Mask-Adapter differs from the heatmap decoder in Deop. Specifically, our Mask-Adapter extracts semantic activation maps from proposal masks, rather than generating heatmaps based on image features and masks, thereby capturing richer contextual and semantic information; (3) during mask-text alignment training, we introduce a mask consistency loss and employ an IoU-based matcher to enhance the model's robustness to varying predicted masks. Additionally, by integrating Mask-Adapter into several well-established methods, we obtain significant improvements and achieve new state-of-the-art results across multiple datasets.

2.2. Adaptation for Vision-Language Models

Recent advancements in vision-language pre-training models, *e.g.*, CLIP [30] and ALIGN [17], have demonstrated strong open-vocabulary recognition capabilities. To further enhance CLIP's performance, various adaptation strategies have been proposed. Notably, prompt tuning methods like CoOp [55] and CoCoOp [54] replace manually crafted prompts with learnable prompts to improve domain-specific adaptation. Feature-based methods, such as CLIP-Adapter [13] and Tip-Adapter [49], introduce lightweight adapters. CLIP-Adapter fine-tunes features through a bottleneck layer, combining learned and pre-trained features, while Tip-Adapter leverages a cache model to enhance few-shot learning with minimal fine-tuning. Despite achieving good results in image classification, CLIP and these adaptations still struggle with dense prediction tasks like open-vocabulary object detection and segmentation.

To address CLIP's limitations in dense prediction tasks, several methods have been proposed. CLIPSelf [39] and SILC [29] leverage self-distillation to align global and local features. EVF-SAM [50] utilizes early vision-language fusion to enable text-prompted SAM for referring expression segmentation tasks. Other approaches, such as MaskCLIP [53], CLIPSurgery [22], SCLIP [37], and GEM [1], modify CLIP's attention structure to enhance its performance on dense prediction tasks. Furthermore, GBA [46] introduces learnable adapters to boost CLIP's generalization in open-vocabulary segmentation, while PixelCLIP [33] improves CLIP's fine-grained pixel representation by leveraging vision foundation models (e.g., DINO [4], SAM [20]) and unlabeled masks. Despite these advancements, CLIP continues to underperform on open-vocabulary region recognition tasks.

3. Approach

3.1. Problem Definition

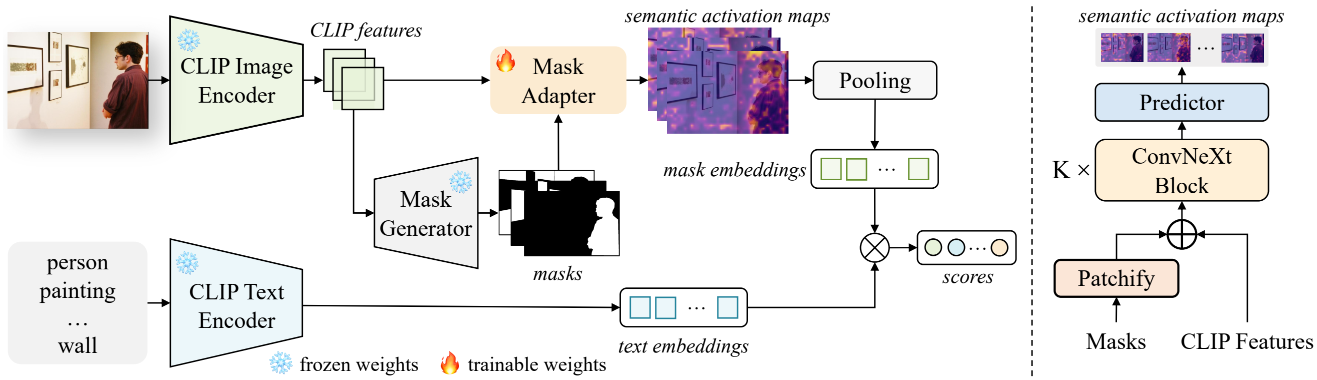
Open-vocabulary segmentation aims to segment images based on arbitrary textual descriptions, rather than a fixed set of predefined classes. This task involves training the model on a set of seen categories $\mathbf{C}_{\text{seen}}$ and evaluating its performance on both the seen categories $\mathbf{C}_{\text{seen}}$ and the unseen categories $\mathbf{C}_{\text{unseen}}$. Existing open-vocabulary segmentation methods [10, 19, 23, 44, 48] are composed of a class-agnostic mask generator and a mask classifier. The mask generator, *e.g.*, Mask2Former [6], produces class-agnostic masks $M_p \in \mathbb{R}^{N\times H\times W}$, and the mask classifier categorizes these masks according to text descriptions.

Given N class-agnostic masks $M_p \in \mathbb{R}^{N\times H\times W}$ generated by the mask generator, the initial approach [10, 44] utilizes *mask cropping* (Fig. 2 (a)), which crops the segmented regions from the image $\mathcal{I}$ and uses *CLIP* to extract mask embeddings, as shown in the following Eq. 1:

$$\mathbf{E}_{\text{mask}} = \textit{CLIP}(\text{crop}(M_p, \mathcal{I})) \in \mathbb{R}^{N\times C}. \tag{1}$$

The mask embeddings are then matched with text embeddings to recognize the segmentation masks. However, the gap between masked and natural images limits the upper bound of this method for open-vocabulary segmentation.

Methods such as [15, 19, 48] employ a *mask pooling* approach (Fig. 2 (b)) to obtain the mask embeddings. In this method, CLIP features are aggregated via mask pooling with the proposal masks, providing greater efficiency



(a) Mask-Adapter for Open-Vocabulary Segmentation

(b) Details of Mask-Adapter

Figure 3. **Overview of Mask-Adapter.** (a) Mask-Adapter for Open-Vocabulary Segmentation. Mask-Adapter can be seamlessly integrated into open-vocabulary segmentation methods based on mask pooling. Mask-Adapter extracts semantic activation maps from CLIP features and proposal masks. Mask embeddings are aggregated according to semantic activation maps, which provide richer contextual and semantic information. The aggregated mask embeddings are then matched with text embeddings to perform mask classification. During training, only the Mask-Adapter is trainable. (b) Details of the Mask-Adapter. After patchifying the masks, masks and CLIP features are fused and processed through ConvNeXt blocks, ultimately outputting the semantic activation maps through a predictor.

than mask cropping, as shown in the following Eq. 2:

$$\mathbf{E}_{\text{mask}} = M_p \cdot \mathcal{F}_{\text{clip}}^T. \tag{2}$$

However, the mask pooling method only captures positional information of the mask region, lacking both contextual and semantic features. This limitation significantly constrains its upper bound in open-vocabulary segmentation.

3.2. Mask-Adapter

Overview. As illustrated in Fig. 3, Mask-Adapter is a simple yet remarkably effective method and can be seamlessly integrated into open-vocabulary segmentation methods to tackle the existing bottlenecks. By feeding masks and CLIP features into Mask-Adapter, it generates semantic activation maps, which highlight informative regions for each mask. With semantic activation maps, we then gather distinctive information from the CLIP features through pooling to obtain mask embeddings. Then mask embeddings are matched with text embeddings to obtain the classification results. During training, we propose a stable mask-text alignment strategy for training. At inference, masks generated by the Mask Generator are fed into Mask-Adapter for mask classification.

Architecture. Fig. 3 illustrates the architecture of Mask-Adapter. Given N class-agnostic masks M_p, we patchify the binary masks into patches via a simple block comprising two strided 3×3 convolutional layers, thus generating mask features $\mathcal{F}_m$ which have the same resolution as the CLIP features. We then fuse CLIP features $\mathcal{F}_{clip}$ and mask features $\mathcal{F}_m$ to obtain the mask region representation $\mathbf{A}_m = \mathcal{F}_m + \mathcal{F}_{clip}$. Next, $\mathbf{A}_m$ is passed through three ConvNeXt [26] blocks to enhance the semantics of the mask regions. The refined features are then processed by a final convolutional layer to generate K semantic activation maps $\mathbf{A}$ for each mask. The semantic activation maps can be expressed as follows:

$$\mathbf{A} = \text{Conv}(\text{ConvNeXt}^3(\mathbf{A}_{\text{m}}))). \tag{3}$$

We then obtain the mask embeddings E_{m} by aggregating information from the CLIP features $\mathcal{F}_{\text{clip}}$ using K semantic activation maps $\bar{\mathbf{A}}$ and taking the average across these embeddings, where $\bar{\mathbf{A}}$ is normalized to 1 for each semantic map. Unlike the mask pooling method, our approach incorporates richer semantic information and includes context beyond the mask region, enhancing the upper bound for open-vocabulary segmentation. The following equation summarizes this process:

$$E_{\text{m}} = \frac{1}{K} \sum_{k=1}^{K} \bar{\mathbf{A}}_k \cdot \mathcal{F}_{\text{clip}}^T, \quad E_{\text{m}} \in \mathbb{R}^{N \times C}. \tag{4}$$

The resulting mask embeddings serve as feature representations for the N masks, which are then matched with text embeddings to classify the segmentation masks.

3.3. IoU-based Matcher

Among the masks generated by the mask generator, multiple masks may correspond to the same object but be misidentified. However, the Hungarian matcher used in previous methods [3, 6, 34], typically finds the optimal match between ground-truth masks and predictions, which fails to capture these misclassified samples. This limitation can lead to overfitting during Mask-Adapter training. To address this, we propose an IoU-based matcher that computes the Intersection over Union (IoU) between each ground-truth mask y_i and predicted mask $\hat{y}_j$, selecting mask pairs based on a predefined IoU threshold. This approach encourages more diverse training samples than one-to-one matching. The IoU-based matcher is defined as:

$$\mathcal{M} = \{(i, j) \mid \mathrm{IoU}(y_i, \hat{y}_j) \geq \mathrm{IoU}_{\text{threshold}}\}, \tag{5}$$

where $\mathcal{M}$ denotes the set of matched pairs. This method enhances the model's robustness by incorporating previously misclassified samples.

3.4. Mask Consistency

As discussed in [14], allocating more space in the feature domain for unseen classes improves the recognition of unseen masks. However, directly using a standard classification loss to train the Mask-Adapter can overly crowd the feature space, limiting the model's ability to recognize unseen categories. In contrast to [14], we adopt a simpler approach by applying a mask consistency loss to mask embeddings of seen class masks. Given the matched predicted and ground-truth masks from the matcher, we feed them into the Mask-Adapter to obtain their respective mask embeddings e^{gt} and e^{pred}. We then compute the cosine similarity loss between these embeddings as follows:

$$\mathcal{L}_{\text{cos}}(e^{\text{gt}}, e^{\text{pred}}) = 1 - \sigma_{\text{cos}}(e^{\text{gt}}, e^{\text{pred}}), \tag{6}$$

where e^{gt} and e^{pred} are the embeddings of the ground-truth and predicted masks, respectively, and σ_{cos} denotes cosine similarity. Mask consistency loss enforces proposal masks with similar IoUs to obtain similar CLIP embeddings, enhancing the recognition of unseen masks and improving the model's robustness.

3.5. Stable Mask-Text Alignment Training

To transfer CLIP's open-vocabulary recognition ability with mask-text alignment training is a natural approach for Mask-Adapter. However, directly using predicted masks leads to overfitting and training instability. To address it, we propose a two-stage training strategy including ground-truth mask warmup and mixed-mask training.

Ground-truth mask warmup. We first train the Mask-Adapter using ground-truth masks to avoid low-quality predicted masks, ensuring stable training. This warmup phase enables the Mask-Adapter to develop strong generalization, making it robust to both ground-truth and predicted masks.

Mixed-mask training. After the warmup training, we mix the predicted masks (from the IoU-based Matcher) and ground-truth masks for training. This mixed-mask training includes low-quality and misclassified masks, enhancing Mask-Adapter's robustness and improving its performance on open-vocabulary segmentation tasks.

Overall training loss. Following [6, 48], we use cross-entropy loss for mask classification and cosine loss to enforce mask consistency. The total loss is:

$$\mathcal{L} = \lambda_{ce} \cdot \mathcal{L}_{ce} + \lambda_{cos} \cdot \mathcal{L}_{cos}, \tag{7}$$

where λ_{ce} and λ_{cos} are the coefficients for the cross-entropy and cosine losses, respectively.

4. Experiments

In this section, we evaluate the performance of Mask-Adapter on open-vocabulary segmentation against the state-of-the-art methods and present detailed ablation studies. Additionally, we explore the potential of integrating Mask-Adapter to SAM [20] for open-vocabulary segmentation.

4.1. Experimental Setup

Baseline. To evaluate Mask-Adapter, we mainly adopt three baseline methods, *i.e.*, FC-CLIP (*open-vocab.*) [48], MAFTP (*open-vocab.*) [19], and Mask2Former (*fixed-vocab.*) [6]. FC-CLIP and MAFTP utilize mask pooling to extract mask embeddings. In addition, we follow [15] and adapt the *fixed-vocabulary* Mask2Former to classify masks with mask pooling for a fair comparison.

Dataset and evaluation metrics. Following the open-vocabulary segmentation setting [9, 23, 32, 44] and baseline models [19, 48], we train our Mask-Adapter separately on COCO-Stuff [2] and on COCO-Panoptic [24]. COCO-Stuff consists of 118k images with 171 categories, while COCO-Panoptic has the same training images but contains 133 categories. Notably, training on COCO-Stuff outperforms COCO-Panoptic on open-vocabulary segmentation tasks [19]. We evaluated our model on widely-used open vocabulary segmentation benchmarks, including ADE20K [52] (A-847 and A-150), Pascal-Context [28] (PC-459 and PC-59), and Pascal-VOC [12] (PAS-20). Additionally, we assess classification performance on ground-truth masks from ADE20K and explore open-vocabulary segmentation by integrating the Mask-Adapter with the Segment Anything Model [20].

We adopt mean Intersection over Union (mIoU) to evaluate semantic segmentation performance, following [9, 10, 23, 32, 44]. To further evaluate the model's generalization ability on unseen categories, we report mIoU^{s} and mIoU^{u} for seen and unseen categories in ablation experiments.

Implementation details. The proposed Mask-Adapter is built on Detectron2 [40] and trained in two phases. In the first phase, we train Mask-Adapter on COCO-Panoptic or COCO-Stuff using ground-truth masks for 20 epochs with a batch size of 8. We follow the training schedule [6, 48] and adopt the AdamW [27] optimizer with an initial learning rate of 1×10^{-4}, weight decay of 0.05, and a multi-step

Method	VLM	Training Dataset	A-150	A-847	PC-59	PC-459	PAS-20
ZegFormer [10]	CLIP ViT-B/16	COCO-Stuff	18.0	5.6	45.5	10.4	89.5
DeOP [16]	CLIP ViT-B/16	COCO-Stuff	22.9	7.1	48.8	9.4	91.7
OvSeg [23]	CLIP ViT-B/16	COCO-Stuff	24.8	7.1	53.3	11.0	92.6
SAN [45]	CLIP ViT-B/16	COCO-Stuff	27.5	10.1	53.8	12.6	94.0
OpenSeg [15]	ALIGN	COCO Panoptic+Loc. Narr.	28.6	8.8	48.2	12.2	72.2
EBSeg [32]	CLIP ViT-B/16	COCO-Stuff	30.0	11.7	56.7	17.3	94.6
SED [41]	CLIP ConvNeXt-B	COCO-Stuff	31.6	11.4	57.3	18.6	94.4
CAT-Seg [9]	CLIP ViT-B/16	COCO-Stuff	31.8	12.0	57.5	**19.0**	94.6
MAFTP* [19]	CLIP ConvNeXt-B	COCO-Stuff	34.5	13.8	57.5	18.5	**95.5**
MAFTP w/ MaskAdapter	CLIP ConvNeXt-B	COCO-Stuff	**35.6**	**14.2**	**58.4**	17.9	95.1
OvSeg [15]	CLIP ViT-L/14	COCO-Stuff	29.6	9.0	55.7	12.4	94.5
ODISE [43]	Stable Diffusion	COCO Panoptic	29.9	11.1	55.3	14.5	-
SAN [45]	CLIP ViT-L/14	COCO-Stuff	32.1	12.4	57.7	15.7	94.6
EBSeg [32]	CLIP ViT-L/14	COCO-Stuff	32.8	13.7	60.2	21.0	96.4
FC-CLIP* [48]	CLIP ConvNeXt-L	COCO Panoptic	34.1	14.8	58.4	18.2	95.4
SED [41]	CLIP ConvNeXt-L	COCO-Stuff	35.3	13.7	60.9	22.1	96.1
GBA [46]	CLIP ConvNeXt-L	COCO Panoptic	35.9	15.1	59.6	18.5	95.8
MAFTP* [19]	CLIP ConvNeXt-L	COCO-Stuff	36.3	15.5	59.5	21.2	96.4
FC-CLIP w/ MaskAdapter	CLIP ConvNeXt-L	COCO Panoptic	36.6	14.1	59.7	19.3	95.5
SMART [8]	CLIP ViT-L/14	COCO Panoptic	36.8	16.1	62.4	23.6	-
MROVSeg [56]	CLIP ViT-L/14	COCO-Stuff	36.9	16.1	**64.1**	**24.1**	**97.6**
CAT-Seg [9]	CLIP ViT-L/14	COCO-Stuff	37.9	16.0	63.3	23.8	97.0
MAFTP w/ MaskAdapter	CLIP ConvNeXt-L	COCO-Stuff	**38.2**	**16.2**	60.4	22.7	95.8

Table 1. **Open Vocabulary Semantic Segmentation Performance.** The best results are indicated in bold. We use mIoU as the evaluation metric. * denotes the results re-evaluated.

Method w/o ensemble	ADE20K mIoU^s	mIoU^u	mIoU
Mask2Former w/ CLIP	34.8	17.5	26.0
Mask2Former w/ Ours	45.3 (+10.5)	23.0 (+5.5)	34.0 (+8.0)
FC-CLIP	34.6	18.6	26.5
FC-CLIP w/ Ours	46.2 (+11.6)	24.8 (+6.2)	35.4 (+8.9)
MAFTP	45.8	27.0	36.3
MAFTP w/ Ours	47.1 (+1.3)	28.5 (+1.5)	37.7 (+1.4)

Table 2. **Results of baseline methods on the representative ADE20K dataset.** To more clearly highlight the effectiveness of our approach, we remove the ensemble methods.

Model	Source	Mask Embedding	Acc
ViT-L/14	Open CLIP	Mask Cropping	29.8
ConvNeXt-L	Open CLIP	Mask Cropping	40.1
ViT-L/14	OVSeg [23]	Mask Cropping	49.9
EVA-L/14[36]	EVA-CLIP	Mask Pooling	32.2
EVA-L/14[36]	CLIPSelf[39]	Mask Pooling	53.1
ConvNeXt-L	Open CLIP	Mask Pooling	53.0
ConvNeXt-L	Open CLIP	**Mask-Adapter**	66.7
ConvNeXt-L	MAFTP	**Mask-Adapter**	**74.1**

Table 3. **Comparison with different mask embedding extraction methods for classification**. We evaluate previous approaches, including Mask Cropping and Mask Pooling methods.

learning rate decay schedule. In the second phase, we employ mixed-mask training on COCO-Panoptic or COCO-Stuff for an additional 10 epochs with a batch size of 16. All models are trained with random flip and a crop size of 1024×1024, while the evaluation is conducted with an image resolution of 896×896. We set the IoU-based Matcher threshold to 0.7 and $\lambda_{ce} = 2.0$, $\lambda_{cos} = 5.0$. The Mask-Adapter generates 16 semantic activation maps for each mask. Following [48], final classification results are obtained by geometric ensembling in-vocabulary classifiers and Mask-Adapter classification results. The training was conducted on 4 NVIDIA RTX 3090 GPUs.

4.2. Main Results

Open-vocabulary segmentation. We follow standard settings [9, 19, 45] to evaluate open-vocabulary segmentation. We classify the models based on the size of the Vision-Language Model (VLM). For base VLM models, Our Mask-Adapter combined with MAFTP [19] achieves the best performance on the A-150, A-847, and PC-59 datasets, with mIoU scores of 35.6, 14.2, and 58.4, respectively, improving mIoU by 1.1, 0.4, and 0.9 compared to the baseline. For large VLMs, our Mask-Adapter with MAFTP outperforms the state-of-the-art method, CAT-

Seg [9], achieving the best results on the challenging A-150 and A-847 datasets. Compared to the baseline FC-CLIP [48], our approach improves mIoU by 2.5, 1.3, and 1.1 on the A-150, PC-59, and PC-459 datasets, respectively, demonstrating its effectiveness.

Comparisons on ADE20K without ensemble strategy. To demonstrate the Mask-Adapter's effectiveness, we conduct experiments on ADE20K without the ensemble strategy. As shown in Tab. 2, incorporating Mask-Adapter significantly boosts the performance of three baselines: Mask2Former (+30.8%), FC-CLIP (+33.5%), and MAFTP (+3.9%), highlighting the substantial improvements in open-vocabulary segmentation.

Comparisons with different mask embedding extraction methods. Tab. 3 presents results of various mask embedding extraction methods on the ADE20K dataset using ground-truth masks. Our method achieves 74.1% accuracy, outperforming the methods in [10, 19, 23, 44, 48] demonstrating the effectiveness of our Mask-Adapter in transferring CLIP's open-vocabulary recognition capabilities to mask classification. Additionally, rows 2, 6, and 7 directly compare mask cropping, mask pooling, and Mask-Adapter using the same model (ConvNeXt-L) and source (OpenCLIP), with Mask-Adapter outperforming the others by 66.3% and 20.5%, respectively.

GT.	Pred.	Warmup	Mask Cons.	ADE20K mIoU^s	ADE20K mIoU^u	ADE20K mIoU
				34.6	18.6	26.5
✓				44.7	23.9	34.1
✓	✓			45.4	23.3	34.3
✓	✓	✓		45.6	23.8	34.6
✓	✓	✓	✓	**46.2**	**24.8**	**35.4**

Table 4. **Ablation study on different training strategies and methods in Mask-Adapter.** GT. refers to ADE20K ground truth masks, Pred. denotes FC-CLIP predicted masks, Mask Cons. indicates whether mask consistency is applied, GT Warmup refers to the warmup using ground truth masks . We remove the ensemble operation here.

4.3. Ablation Experiments

Ablation study on Mask-Adapter. As shown in Tab. 4, replacing Mask Pooling with the ground-truth trained Mask-Adapter in FC-CLIP improves mIoU by 10.1 and 5.3 for seen and unseen categories, respectively. However, directly using mixed predicted and ground-truth masks for training slightly reduces performance on unseen classes. Introducing Ground-truth warmup and mask consistency loss improves mIoU by 0.3 and 0.8, respectively, highlighting the importance of Ground-Truth Warmup and Mask Consistency in enhancing the model's robustness.

Matcher	Threshold	ADE20K mIoU^s	ADE20K mIoU^u	ADE20K mIoU
Hungarian	-	44.4	22.3	33.2
IoU-based	0.7	**45.3**	**23.0**	**34.0**
IoU-based	0.8	45.0	22.8	33.7

Table 5. **Ablation study of different matchers and IoU thresholds.** We evaluate the impact of various matchers and IoU thresholds on the ADE20K dataset. We adopt Mask2Former as baseline here.

Ablation study of different matchers and IoU thresholds. As shown in Tab. 5, we evaluate the impact of different matchers on the ADE20K dataset. With an IoU-based matcher threshold of 0.7, the model improves by 0.8 mIoU overall, and 0.9 mIoU for seen and 0.7 mIoU for unseen classes.

	Patch Embedding	GT mIoU	FC-CLIP mIoU
(a)	Concat	42.5	35.2
(a)	Add + 2 × 3 × 3 Conv	**43.4**	**35.4**
(a)	Add + 4 × 4 Conv	**43.4**	35.2
	Num Blocks	**GT mIoU**	**FC-CLIP mIoU**
(b)	2	43.1	**35.4**
(b)	3	43.4	**35.4**
(b)	4	**43.7**	35.3
	CLIP Feature	**GT mIoU**	**FC-CLIP mIoU**
(c)	with	**43.4**	**35.4**
(c)	w/o	35.5	32.5

Table 6. **Ablation study on different designs in Mask-Adapter**, analyzing the impact of different patch embedding methods, ConvNeXt block counts, and input CLIP feature usage. Section (a) compares patch embedding strategies, (b) examines the effect of varying ConvNeXt block numbers, and (c) highlights the importance of including CLIP features. The experiments are conducted during the warm-up training phase.

Ablation study on different designs in Mask-Adapter. Tab. 6 presents an ablation study on key Mask-Adapter design choices. In (a), we show that adding CLIP features to mask features outperforms simple concatenation, proving this fusion strategy is the most effective. Our mask patchify approach, with two 3×3 convolutions, outperforms a single 4×4 convolution. In Tab. 6 (b), we find that increasing ConvNeXt block count improves mask classification accuracy with ground-truth masks, but too many blocks cause overfitting. Three blocks strike the optimal balance. In Tab. 6 (c),

Figure 4. **Visualizations of Semantic Activation Maps.** We present visualizations of the semantic activation maps and their corresponding segmentation masks. For each input image, the top row shows the semantic activation maps, while the bottom row displays the segmentation masks. The semantic activation maps emphasize the most discriminative regions of the mask. Best viewed on screen after zooming in.

incorporating CLIP features significantly improves Mask Accuracy and mIoU, enhancing semantic representation.

Ablation of ground-truth masks warmup datasets. Tab. 7 evaluates the impact of different dataset combinations for ground-truth warmup training in open-vocabulary segmentation. While COCO-Panoptic is a standard choice, we observed overfitting with this dataset, leading to reduced accuracy for unseen categories over time. To address this, we incorporate additional datasets. LVIS enhances Mask-Adapter's robustness, likely due to its richer textual information. However, integrating a subset of the Grand dataset [31] (2% of the SA-1B [20]) led to performance degradation compared to the COCO-only baseline, likely due to mask quality problems.

Datasets	GT mIoU	FC-CLIP mIoU
COCO-P	43.4	35.4
COCO-P, LVIS	**44.5**	**36.4**
COCO-P, LVIS, Grand	43.6	35.1

Table 7. **Ablations on ground-truth warmup training data.** We evaluates the impact of different dataset combinations for ground-truth warmup training in open-vocabulary segmentation.

4.4. Segment Anything with Mask-Adapter

Mask-Adapter can be seamlessly integrated into SAM without training, significantly enhancing its performance on open-vocabulary segmentation tasks. Following [18], we conduct experiments under *open-vocabulary settings*. Tab. 8 shows that SAM with the Mask-Adapter outperforms SAM w/ CLIP or w/ MAFT across multiple datasets, highlighting the Mask-Adapter's effectiveness in improving the classification accuracy of masks generated by SAM.

	A-847	A-150	PC-459	PC-59
SAM w/ CLIP	7.1	17.9	6.4	34.4
SAM w/ MAFT	10.1	29.1	12.8	**53.5**
SAM w/ Mask-Adapter	**11.4**	**31.5**	**16.8**	49.5

Table 8. **Comparison of SAM and SAM with Mask-Adapter under open-vocabulary settings.** Following [18], we evaluate performance using SAM-H with a standard CLIP model and MAFT [18], as well as our Mask-Adapter approach.

5. Conclusion

In this work, we reveal severe limitations of prevalent methods using mask pooling and predicted masks for open-vocabulary segmentation and present a simple yet effective Mask-Adapter. Instead of directly using proposal masks, Mask-Adapter extracts semantic activation maps from masks and obtains mask embeddings by aggregating CLIP features according to the semantic and contextual information provided by the semantic activation maps for mask classification. We integrate Mask-Adapter into several well-established open-vocabulary segmentation methods, demonstrating significant performance gains and achieving state-of-the-art results across multiple datasets. Mask-Adapter also extends to SAM without training and achieves impressive results on several open-vocabulary segmentation benchmarks.

Acknowledgement. This work was in part supported by National Science and Technology Major Project under Grant No. 2023YFF0905400.

References

[1] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 3

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5

[3] 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. 4

[4] 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. 3

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 1

[6] 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. 3, 4, 5

[7] Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4433–4442, 2022. 1, 2

[8] Yong Xien Chng, Xuchong Qiu, Yizeng Han, Kai Ding, Wan Ding, and Gao Huang. Semantic refocused tuning for open-vocabulary panoptic segmentation. *arXiv preprint arXiv:2409.16278*, 2024. 3, 6

[9] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Catseg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2, 5, 6, 7

[10] 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. 1, 2, 3, 5, 6, 7

[11] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 3

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 3

[14] Jiannan Ge, Lingxi Xie, Hongtao Xie, Pandeng Li, Xiaopeng Zhang, Yongdong Zhang, and Qi Tian. Alignzeg: Mitigating objective misalignment for zero-shot semantic segmentation. In *European Conference on Computer Vision*, pages 142–161. Springer, 2025. 5

[15] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2, 3, 5, 6

[16] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1086–1096, 2023. 1, 2, 3, 6

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 3

[18] 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. 3, 8

[19] Siyu Jiao, Hongguang Zhu, Jiannan Huang, Yao Zhao, Yunchao Wei, and Humphrey Shi. Collaborative vision-text representation optimizing for open-vocabulary segmentation. *arXiv preprint arXiv:2408.00744*, 2024. 1, 2, 3, 5, 6, 7

[20] 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. 3, 5, 8

[21] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022. 2

[22] 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. 3

[23] 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. 1, 2, 3, 5, 6, 7

[24] 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. 5

[25] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2(1):16, 2024. 2

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[27] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 5

[29] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. *arXiv preprint arXiv:2310.13355*, 2023. 3

[30] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3

[31] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 8

[32] Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, and Changxin Gao. Open-vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28412–28421, 2024. 1, 2, 5, 6

[33] Heeseong Shin, Chaehyun Kim, Sunghwan Hong, Seokju Cho, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Towards open-vocabulary semantic segmentation without semantic labels. *arXiv preprint arXiv:2409.19846*, 2024. 3

[34] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016. 4

[35] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021. 1

[36] 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. 6

[37] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. 3

[38] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 1

[39] 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. *arXiv preprint arXiv:2310.01403*, 2023. 3, 6

[40] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 5

[41] 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. 2, 6

[42] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021. 1

[43] 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. 2, 6

[44] 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. 1, 2, 3, 5, 7

[45] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3, 6

[46] Wenhao Xu, Changwei Wang, Xuxiang Feng, Rongtao Xu, Longzhao Huang, Zherui Zhang, Li Guo, and Shibiao Xu. Generalization boosted adapter for open-vocabulary segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1, 2, 3, 6

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

[48] 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*, 36:32215–32234, 2023. 1, 2, 3, 5, 6, 7

[49] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3

[50] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for

text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 3

[51] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 1

[52] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 1, 5

[53] 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. 3

[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 3

[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3

[56] Yuanbing Zhu, Bingke Zhu, Zhen Chen, Huan Xu, Ming Tang, and Jinqiao Wang. Mrovseg: Breaking the resolution curse of vision-language models in open-vocabulary semantic segmentation. *arXiv preprint arXiv:2408.14776*, 2024. 6