Collaborative Vision-Text Representation Optimizing for Open-Vocabulary Segmentation

Siyu Jiao[1,2⋆], Hongguang Zhu[1,2⋆], Jiannan Huang[1,3], Yao Zhao[1,2], Yunchao Wei[1,2], and Humphrey Shi[3,4]

[1] Institute of Information Science, Beijing Jiaotong University
[2] Peng Cheng Laboratory
[3] Georgia Institute of Technology
[4] Picsart AI Research (PAIR)
cici579869@gmail.com

Abstract. Pre-trained vision-language models, *e.g.* CLIP, have been increasingly used to address the challenging Open-Vocabulary Segmentation (OVS) task, benefiting from their well-aligned vision-text embedding space. Typical solutions involve either freezing CLIP during training to unilaterally maintain its zero-shot capability, or fine-tuning CLIP vision encoder to achieve perceptual sensitivity to local regions. However, few of them incorporate vision-text collaborative optimization. Based on this, we propose the Content-Dependent Transfer to adaptively enhance each text embedding by interacting with the input image, which presents a parameter-efficient way to optimize the text representation. Besides, we additionally introduce a Representation Compensation strategy, reviewing the original CLIP-V representation as compensation to maintain the zero-shot capability of CLIP. In this way, the vision and text representation of CLIP are optimized collaboratively, enhancing the alignment of the vision-text feature space. To the best of our knowledge, we are the first to establish the collaborative vision-text optimizing mechanism within the OVS field. Extensive experiments demonstrate our method achieves superior performance on popular OVS benchmarks. In open-vocabulary semantic segmentation, our method outperforms the previous state-of-the-art approaches by +0.5, +2.3, +3.4, +0.4 and +1.1 mIoU, respectively on A-847, A-150, PC-459, PC-59 and PAS-20. Furthermore, in a panoptic setting on ADE20K, we achieve the performance of 27.1 PQ, 73.5 SQ, and 32.9 RQ. Code will be available at MAFT-Plus.

Keywords: Open-Vocabulary Segmentation · Fine-tuning

1 Introduction

Segmentation stands as the most popular basic topics in computer vision, traditional segmentation models [4, 10, 14, 15, 43] are only capable of segmenting a few predefined categories within a closed vocabulary [3, 9], notably smaller

⋆ Equal contribution

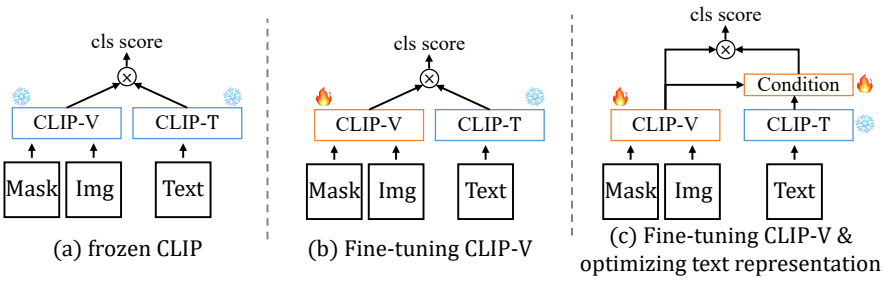

Fig. 1: Different learning frameworks for open-vocabulary segmentation, from the perspective of whether to freeze CLIP. (a) The "frozen CLIP" paradigm. [22, 25, 36, 37] (b) Fine-tuning CLIP-V [18]. (c) Our MAFT+ framework enables to optimize both CLIP-V and CLIP-T.

than the human-used categories for describing the real world. Therefore, open-vocabulary segmentation (OVS) [2, 12, 13, 31] is introduced to segment objects using arbitrary categories described by texts.

Recently, large-scale visual-language pre-training models (*e.g.* CLIP [26] and ALIGN [17]) learn representation with cross-modal alignment and show strong zero-shot capability, leading to the increased adoption for tackling the challenging OVS task [8, 22, 25, 36]. A mainstream solution follows the "decoupling" paradigm, which executes the open-vocabulary segmentation with two steps: 1) employing a Proposal Generator to produce class-agnostic mask proposals and 2) leveraging a pre-trained CLIP to classify each mask proposal via similarity matching in the aligned image-text feature space. The above-mentioned paradigm can be categorized into two groups hinges on whether CLIP is frozen during the training process, as depicted in Fig. 1a, b.

In order to retain the strong zero-shot capability of CLIP when classifying mask proposals, most previous works [22, 25, 36, 37] choose to freeze the pre-trained CLIP model (Fig. 1a). They execute with either masked-crops or masked-attention, when processing images and masks within CLIP-V. Considering the domain gap between image-level pre-training of CLIP and pixel-level application of segmentation, these approaches compromise the representational ability of CLIP, and fail to fit the distribution of segmentation tasks well. Recent work MAFT [18] highlights the frozen CLIP is insensitive to different mask proposals and often yields similar predictions. It designs a mask-aware fine-tuning strategy to enhance the sensitivity of CLIP-V to local regions (Fig. 1b). While MAFT partially addresses the insensitivity issue, it comes with some new problems: 1) only updating CLIP-V constrains the overall optimization space, thereby limiting the alignment of vision and text representation. 2) fine-tuning CLIP-V on downstream datasets leads to the degradation of generalization ability.

To address the aforementioned problems, we introduce a collaborative Vision-Text representation fine-tuning framework as the enhanced version of MAFT, named MAFT+. As shown in Fig. 1c. Specific to enhance the alignment of vision-

text representation, we incorporate CLIP-T into the fine-tuning process to concurrently optimize the text representation. This vision-text joint optimization alleviates the training complexity and enhances the vision and text alignment. Considering the challenging GPU memory requirements for fine-tuning CLIP-T, we introduce a Content-Dependent Transfer (CDT) following CLIP-T to optimize text representation in a parameter-efficient way. CDT utilizes Transformer Layers to condition text embeddings on each input image rather than fixed once generated by CLIP-T, mitigating the computational burden while preserving the effectiveness of the fine-tuning process. Moreover, to maintain the zero-shot capality during CLIP-V fine-tuning, we draw inspiration from preventing Catastrophic Forgetting [23] in continual learning, and devise a Representation Compensation (RC) strategy. This strategy aims to preserve CLIP's zero-shot capability by reviewing the pre-trained representation of an original CLIP-V as a form of compensation.

Overall, our contributions are summarized as follows:

- Our MAFT+ represents the first collaborative framework to jointly optimize vision-text representation in OVS. This collaborative design mitigates training complexity and enhances alignment in the vision-text feature space.
- The Content-Dependent Transfer is proposed to unleash the optimization potential of CLIP-T through parameter-efficient fine-tuning. The Representation Compensation achieves effective CLIP-V fine-tuning while maintaining the original zero-shot capability.

We evaluate our MAFT+ on the commonly used open-vocabulary *semantic* and *panoptic* segmentation benchmarks: Pacal-Context [24], Pascal-VOC [9], and ADE20K [45]. Compared with the prior open-vocabulary *semantic* results, MAFT+ enhances the performance of A-847 [45], A-150 [45], PC-459 [24], PC-59 [24] and PAS-20 [9] datasets by +0.5, +2.3, +3.4, +0.4 and +1.1 mIoU respectively. Furthermore, we conduct experiments in a *panoptic* setting, where MAFT+ achieves the performance of 27.1 PQ, 73.5 SQ, and 32.9 RQ on the ADE20K dataset. Notably, our approach outperforms the existing OVS methods and establishes new state-of-the-art results across all evaluated datasets.

2 Related Work

Open-Vocabulary Segmentation [28] is established to break category restrictions and perform segmentation across arbitrary categories. Earlier works [2, 12, 21, 31, 33] use large pre-trained vision-language models to perform open-vocabulary segmentation, they leverage rich alignment features from image-text pairs. Recent approaches [5, 8, 11, 18, 22, 25, 34–37] decouple the open-vocabulary segmentation into mask proposals generation and mask proposals classification, they first generate a series of mask proposals and then utilize CLIP [26] or ALIGN [17] for classification. Specifically, Zegformer [8] first uses mask&crop to get sub-images based on mask proposals, feeding them into CLIP for mask classification. The following approaches ZSSeg [36] and OVSeg [22], train CLIP adapters to boost performance. In order to improve the classification ability

of the vision-language models, OpenSeg [11] takes extra image-caption pairs to scale up training data. FreeSeg [25] unifies semantic, instance, and panoptic tasks and performs fusion training. ODISE [34] utilizes a strong text-to-image diffusion model [27] to obtain a well-aligned image-text feature space. SAN [35] and FC-CLIP [37] design the end-to-end frameworks by exploiting a single frozen CLIP as the backbone. Recently, MAFT [18] introduces a CLIP-V fine-tuning strategy, allowing CLIP-V to be sensitive to different mask proposals.

Pre-trained model fine-tuning is widely used for fitting the distribution to downstream tasks. Specific to segmentation, traditional close-set methods [4, 14, 15, 43] typically use a lower learning rate (*e.g.* $\frac{1}{10}$) to fine-tune the image encoder, transferring pre-trained knowledge to segmentation tasks. However, this strategy may be suboptimal for data-limited scenarios such as few-shot segmentation, zero-shot segmentation and incremental segmentation due to the daunting *overfitting* problem. To tackle this, SVF [29] fine-tunes only a subset of parameters in the pre-trained image encoder, adapting pre-trained knowledge to few-shot segmentation. [22] applies prompt-tuning to learn image prompts using annotated data, adapting CLIP-V to masked images. Some continual segmentation approaches utilize techniques like contrastive learning [40–42], distillation [38] and EMA [32] to avoid catastrophic forgetting.

In a recent development, MAFT [18] conducts a mask-aware CLIP fine-tuning strategy by aligning CLIP's classification score with the IoU score. Although this approach partially adapts CLIP-V to segmentation tasks, it exclusively optimizes CLIP-V representation, potentially amplifying the training difficulty and risking overfitting on fixed text embeddings. This observation motivates our exploration of collaborative optimization strategies for both vision and text representation.

3 Preliminary

Problem Setting. Open-vocabulary segmentation addresses the task of training a segmentation model capable of segmenting arbitrary objects using text descriptions. Given two category sets C_{train} and C_{test}, where C_{train} and C_{test} are unequal in terms of object categories ($C_{train} \neq C_{test}$). The model is trained on C_{train} and directly tested on C_{test}. Typically, C_{train} and C_{test} are described by noun words (*e.g.* sky, sea, mount...).

***mask-aware* Loss Function.** [18] proposes a *mask-aware* loss ($\mathcal{L}_{ma}$) to fine-tune CLIP-V for sensitivity to local regions. The primary objective of $\mathcal{L}_{ma}$ is to assign high classification scores to high-quality proposals and low scores to low-quality proposals. This is achieved by utilizing the Intersection over Union (IoU) score S^{IoU} derived from ground-truth as supervision and aligning it with the CLIP classification score S^{cls} to induce mask awareness. The *mask-aware* loss is calculated using the SmoothL1 function:

$$\mathcal{L}_{ma} = \text{SmoothL1}(S^{cls}, S^{IoU}) \tag{1}$$

In this paper, we use $\mathcal{L}_{ma}$ to fit the distribution of CLIP with OVS. Furthermore, we delve into CLIP fine-tuning techniques, and propose a novel CLIP fine-tuning strategy by collaboratively optimizing the distribution of CLIP-V and CLIP-T.

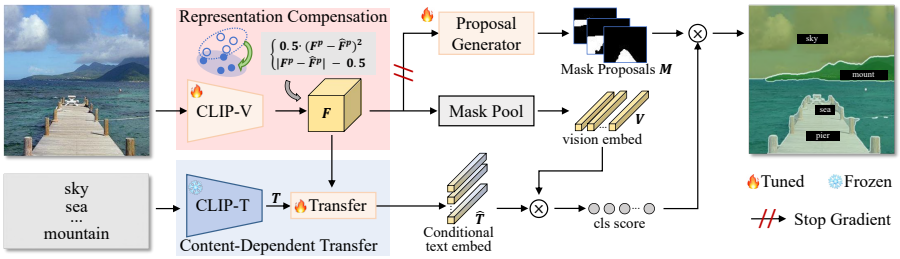

Fig. 2: Overview of the MAFT+. We use CLIP-V as the backbone to extract image features. A Proposal Generator is trained to generate mask proposals. The Representation Compensation strategy reviews the vision representation to preserve the zero-shot capability of CLIP (red part); the Content-Dependent Transfer enables the text embeddings conditioned on input image, and achieves text representation optimizing in a parameter-efficient fine-tuning way. (blue part).

4 Methodology

We introduce MAFT+, a method for collaboratively optimizing CLIP's vision and text representation. The complete framework of the MAFT+ is shown in Fig. 2, we use the Convnext-Large CLIP model for illustration. Within MAFT+, CLIP-V serves as the vision backbone, and a Proposal Generator is trained to generate class-agnostic mask proposals (Sec. 4.1). Simultaneously, the representation of CLIP-V and CLIP-T is collaboratively optimized. We introduce the Representation Compensation (RC) strategy for CLIP-V fine-tuning (Sec. 4.2), and propose the Content-Dependent Transfer (CDT) for parameter-efficient CLIP-T fine-tuning (Sec. 4.3). Finally, we outline the loss functions in Sec. 4.4.

4.1 Feature Extraction & Proposal Generator

Feature Extraction. We utilize a pre-trained convolutional CLIP-V for extracting features from an input image I. Denoting each stage of CLIP-V's output as $F = \{F^i\}, i \in [0, 1, 2, 3]$. F^0, F^1, F^2, F^3 have strides of {4, 8, 16, 32} with respect to the input image.

Proposal Generator. We follow the common design [8, 18, 22, 25, 36, 37, 39] to use MaskFormer [6, 7] as the Proposal Generator. Since the Hungarian matching [20] is used in the training process, only a subset of the mask proposals is optimized. This matching strategy enhances generalizability of the Proposal Generator, ensuring it segment masks of novel categories. Given the image features F, the Proposal Generator generates a set of N mask proposals $M = \{m_i\}_{i=1}^{N} \in \mathbb{R}^{N \times H \times W}$.

During the training process, we stop the gradient flow from CLIP-V to the Proposal Generator. This measure is taken to avoid the potential overfitting of CLIP-V on the training categories.

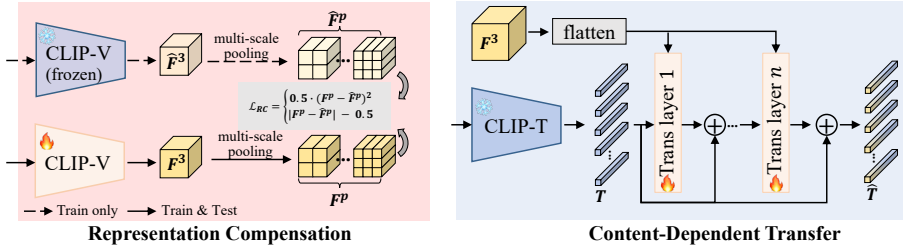

Fig. 3: Details of Representation Compensation and Content-Dependent Transfer.

4.2 Representation Compensation

The representation Compensation (RC) strategy aims to review the original representation of CLIP as compensation during the training phase. Details of Representation Compensation are shown in Fig. 3. Within RC, we use a frozen CLIP-V (denoted as CLIP-V*) to generate the original CLIP-V features during training. Extracting the last stage output from the CLIP-V* ($\hat{F}^3$) and the fine-tuned CLIP-V (F^3), $\hat{F}^3$ and F^3 are expected to be similar to avoid Catastrophic Forgetting. However, direct per-pixel alignment is not feasible, as it would result in the loss of region-level differences. Therefore, we devise multiple grids of average pooling (AvgPooling) to generate multi-scale features, and ensure the consistency of the features after pooling.

Given an arbitrary feature $f \in \mathbb{R}^{d\times h\times w}$, an AvgPooling operation with grid size of $k \times k$ can be formulated as:

$$f^{pool} = \text{AvgPooling}(f, k), f^{pool} \in \mathbb{R}^{d\times k\times k}. \tag{2}$$

In our default design, we use AvgPooling with $K = \{1, 2, 4\}$ to perform pooling $\hat{F}^3$ and F^3 into $\{1\times1, 2\times2, 4\times4\}$ grids, denoting as $\hat{F}^p$ and F^p. Specifically, $\hat{F}^p = \text{AvgPooling}(\hat{F}^3, K)$ and $F^p = \text{AvgPooling}(F^3, K)$. Then, we use SmoothL1 Loss to minimize the difference as follows:

$$\mathcal{L}_{rc} = \text{SmoothL1}(F^p, \hat{F}^p), \tag{3}$$

$$\text{SmoothL1}(F^p, \hat{F}^p) = \begin{cases} 0.5 \cdot (F^p - \hat{F}^p)^2, & \text{if } |F^p - \hat{F}^p| < 1 \\ |F^p - \hat{F}^p| - 0.5, & \text{otherwise} \end{cases} \tag{4}$$

With RC to compensate F^3 original CLIP's representation, the CLIP-V maintains the zero-shot capability during fine-tuning. We apply Mask Pooling [37] on the F^3 to generate vision embeddings ($V \in \mathbb{R}^{N\times d}$) for each mask proposal.

4.3 Content-Dependent Transfer

Given a set of class names $C = \{C_1, C_2...C_n\}$, we use the predefined templates [18, 34–37] to generate sentences corresponding to these class names, *e.g.*, "a

photo of a $\{C_i\}$; There is a $\{C_i\}$ in the scene...", these sentences are then fed into CLIP-T to generate embeddings of each sentence. The embeddings of the same classes are averaged to obtain text embedding ($T \in \mathbb{R}^{d\times|C|}$). d is the dimension of the embedding, and $|C|$ is the number of class names.

To optimize CLIP-T representation T, we propose the Content-Dependent Transfer (CDT), which involves a sequence of Transformer Layers performing cross-attention with vision feature F^3. Details of the CDT are illustrated in Fig. 3. We take the last stage feature of CLIP-V (F^3) and the text embeddings T as the inputs for CDT. F^3 is first Flatten at spatial dimension, denoted as $F^3_{flat} \in \mathbb{R}^{d\times hw}$. Then, we use n sequential Transformer Layers to process T and F^3_{flat}, while incorporating a shortcut connection. This process can be formulated as:

$$T_{i+1} = \text{TransLayer}_\text{i}(T_i, F^3_{flat}) + T_i, \quad i = 1, 2...l. \tag{5}$$

In our default setting, l is set to 2. The resulting output of the CDT is denoted as the conditioned text embeddings ($\hat{T}$). Specifically,

$$\text{TransLayer}(a, b) = \text{Softmax}(\frac{\text{Que}(a)\cdot\text{Key}(b)}{\sqrt{d}})\cdot\text{Val}(b), \tag{6}$$

where Que$(\cdot)$, Key$(\cdot)$, and Val$(\cdot)$ represent linear projections, d is the dimension of the input vectors, we assume all vectors have the same dimension d by default. In Eq. 6, we simplify the expression of Multihead Attention and LayerNorm in Transformer. Note that the CLIP-T remains frozen during training, and only the Transformer Layers are trained to optimize the CLIP-T representation. Therefore, the parameter-efficient CLIP-T fine-tuning is established, with $\hat{T}$ is conditioned on the input images.

We investigate various designs to optimize the CLIP-T representation (T), including fine-tuning CLIP-T, training an additional MLP, incorporating description guidance, etc. Further details are presented in Sec. 5.3.

4.4 Objective

After getting the conditional text embeddings $\hat{T}$, we perform matrix multiplication on $\hat{T}$ and V to derive the classification score S^{cls} for the mask proposals. Subsequently, we multiply S^{cls} with M to obtain the final output.

We use the *mask-aware* loss [18] ($\mathcal{L}_{ma}$, Eq. 1) on S^{cls} to optimize the representation of both CLIP-V and CLIP-T. Considering the $\mathcal{L}_{ma}$ may induce overfitting on the training categories and reduce the transferability of CLIP, we introduce $\mathcal{L}_{rc}$ (Sec. 4.2) to compensate CLIP's representation during training. Meanwhile, we follow Mask2Former [6] to adopt the same loss functions ($\mathcal{L}_P$) to train the Proposal Generator without any special design. Therefore, the final loss function ($\mathcal{L}$) can be formulated as: $\mathcal{L} = \mathcal{L}_P + \lambda_1\mathcal{L}_{ma} + \lambda_2\mathcal{L}_{rc}$, where $\lambda_1 = 1$ and $\lambda_2 = 0.1$.

Note that we stop the gradient from CLIP-V to Proposal Generator. The CLIP-V is not optimized by $\mathcal{L}_P$.

Modifications in the panoptic setting. The $\mathcal{L}_{ma}$ is tailored for semantic segmentation and lacks the ability to capture instance-level information. We explore

adapting the $\mathcal{L}_{ma}$ to panoptic segmentation with the following modification. Specifically, when a mask contains multiple instances, we use binary ground-truth (GT) to mask out redundant instances, retaining only the instance with the highest IoU score with GT. This change allows CLIP-V to learn instance-level knowledge, making $\mathcal{L}_{ma}$ applicable to panoptic segmentation.

5 Experiments

5.1 Setting

Dataset. We conduct experiments on popular open-vocabulary segmentation benchmarks, including COCO-Stuff, COCO-Panoptic, Pascal-VOC, Pascal-Context and ADE20K. We train MAFT+ on COCO-Stuff and testing on ADE20K (A-847, A-150), Pascal-Context (PC-459, PC-59), and Pascal-VOC (PAS-20) to evaluate the performance of open-vocabulary *semantic* segmentation. Then, we evaluate MAFT+ in open-vocabulary *panoptic* settings [5, 34, 37], *i.e.*, training on COCO-Panoptic and testing on ADE20K.
More details of the dataset settings are provided in the *Appendix*.
Evaluation Metrics. To quantitatively evaluate the performance, we follow standard practice [8, 22, 34–37]. Semantic segmentation results are evaluated with mean Intersection over Union (mIoU) [9]. Panoptic segmentation results are evaluated with the panoptic quality (PQ), segmentation quality (SQ) and recognition quality (RQ) [19].
Implementation details. We employ ConvNeXt-Large CLIP from OpenCLIP [16]. The Proposal Generator is built following the default settings of Mask2Former [6]. We set the number of class-agnostic mask proposals to 100 ($N = 100$). During training, the model is optimized with AdamW optimizer with a weight-decay of 0.05. The learning rate is set to 1×10^{-5} for CLIP-V and 1×10^{-4} for other modules. We use a crop size of 1024×1024. The model is trained for 60,000 iterations on COCO with 4 NVIDIA A100 GPUs.

5.2 Comparisons with State-of-the-art Methods

In this section, we compare our proposed MAFT+ with the state-of-the-art open-vocabulary *semantic* segmentation methods and open-vocabulary *panoptic* segmentation methods.
Comparisons in the *semantic* setting. In Tab. 1, we present the performance of MAFT+ on various benchmarks. MAFT+ demonstrates a significant improvement over existing open-vocabulary segmentation models, achieving a performance boost of +0.5, +2.3, +3.4, +0.4, +1.1 mIoU across A-847, PC-459, A-150, PC-59, and PAS-20, respectively. Moreover, compared to MAFT [18], our MAFT+ eliminates the need for an additional fine-tuned CLIP-V. MAFT+ applies an end-to-end pipeline, facilitating both the training and testing processes.
Comparisons in the *panoptic* setting. In Tab. 2, we evaluate our MAFT+ on ADE20K, the main evaluation dataset of open-vocabulary panoptic segmentation. With the aforementioned modifications, our approach achieves new state-of-the-art performance. Compared to FC-CLIP without the ensemble strategy

Table 1: Open-vocabulary *semantic* segmentation performance. mIoU is used to evaluate the performance. * denotes additional ensemble operation [37] used during testing.

	VLM	**A-847**	**A-150**	**PC-459**	**PC-59**	**PAS-20**
OpenSeg [ECCV22] [11]	ALIGN	8.8	28.6	12.2	48.2	72.2
OVSeg [CVPR23] [22]	ViT-L	9.0	29.6	12.4	55.7	94.5
SAN [CVPR23] [35]	ViT-L	12.4	32.1	15.7	57.7	94.6
ODISE [CVPR23] [34]	ViT-L	11.1	29.9	14.5	57.3	-
FC-CLIP [NeurIPS23] [37]	ConvNeXt-L	11.2	26.6	12.7	42.4	89.5
FC-CLIP* [NeurIPS23] [37]	ConvNeXt-L	14.8	34.0	18.2	58.4	95.4
MAFT [NeurIPS23] [18]	ViT-L	12.7	33.0	16.2	59.0	92.1
MAFT [NeurIPS23] [18]	ConvNeXt-L	13.1	34.4	17.0	57.5	93.0
MAFT+ (ours)	ConvNeXt-L	15.1	36.1	21.6	59.4	96.5

Table 2: Open-vocabulary *panoptic* segmentation performance on ADE20K. PQ, SQ, and RQ are used for evaluation. The best results are highlighted with red.

	PQ	SQ	RQ
FreeSeg [CVPR22] [25]	16.3	-	-
ODISE [CVPR22] [34]	22.6	-	-
MaskCLIP [ICML23] [44]	15.1	70.4	19.2
OPSNet [ICCV23] [5]	19.0	52.4	23.0
FC-CLIP [NeurIPS23] [37]	21.9	71.5	26.4
FC-CLIP* [NeurIPS23] [37]	26.8	71.5	32.2
MAFT+ (ours)	27.1	73.5	32.9

(3^{rd} last results), our MAFT+ outperforms it by +5.2 PQ, +2.0 SQ and +6.5 RQ. Although the ensemble strategy greatly improves FC-CLIP's performance, our model still outperforms FC-CLIP* across all evaluation metrics.

Analysis of the ensemble strategy in FC-CLIP. FC-CLIP ensembles the classification score of Mask2Former and CLIP, along with two hyper-parameters to balance these scores. As shown in Tab. 1 and Tab. 2, the ensemble operation significantly improves FC-CLIP's performance. *i.e.*, 42.4→58.4 mIoU on PC-59, 21.9→26.8 PQ on ADE20K. However, **this improvement stems from the overlap of categories between training and testing datasets**. Moreover, determining the two critical hyper-parameters requires numerous repeated experiments. Based on this, in our default settings, we remove this ensemble operation, and solely use the CLIP for classification.

5.3 Ablation Study

We conduct ablation studies on various choices of designs of our MAFT+, and showcase their contribution to the final results in Tab. 3, 4, 5. We freeze the CLIP-V and remove the Content-Dependent Transfer as the baseline model (*i.e.* representation of a frozen CLIP).

Component-wise ablations. To understand the effect of each component in the MAFT+, including the Representation Compensation (RC) strategy and

Table 3: Ablation on components of **MAFT**+. Here RC and CDT denote Representation Compensation and Content-Dependent Transfer. Note that "tune CLIP-T" represents optimizing the distribution of text-embeds, not directly fine-tuning CLIP-T.

	A-847	A-150	PC-459	PC-59	PAS-20
frozen CLIP (baseline)	11.2	26.6	12.7	42.4	89.5
+ CDT (tune CLIP-T)	13.3 +2.1	32.4 +5.8	17.2 +4.5	55.2 +12.8	94.7 +5.2
+ RC (tune CLIP-V)	14.6 +3.4	34.8 +8.2	18.2 +5.5	57.1 +14.7	95.3 +5.8
+ CDT & RC	15.1 +3.9	36.1 +9.5	21.6 +8.9	59.4 +17.0	96.5 +7.0

the Content-Dependent Transfer (CDT). We start with a frozen CLIP as the baseline model, and gradually add each design. (Tab. 3). The frozen CLIP yields inferior performance due to CLIP's region-unaware property (1^{st} row). Then, Content-Dependent Transfer optimizes CLIP Text representation and promotes the alignment of vision and text embeddings, resulting in an improvement of +5.8 mIoU on A-150 and +12.8 mIoU on PC-59 (2^{nd} row). Using only Representation Compensation for fine-tuning CLIP-V produces decent performance (the 3^{rd} result), 26.6→34.8 on A-150, 42.4→57.1 on PC-59 in terms of mIoU. Finally, introducing CDT and RC collaboratively learns effective vision and text alignment representation, fitting the distribution of CLIP from image-level to segmentation tasks, further enhancing the performance to establish state-of-the-art benchmarks. (last row).

Effect of Content-Dependent Transfer. Optimizing CLIP text representation is an essential design of MAFT+. We investigate various designs to optimize the CLIP-T representation in Fig. 4, including direct fine-tuning of CLIP-T parameters, training with additional MLP, training with class-description sentences by GPT, and training with class-description embeddings by Llama-2. Tab. 4 presents the results of different designs for optimizing CLIP text representation. Here, we remove Representation Compensation strategy, and keep the CLIP-V frozen for analysis.

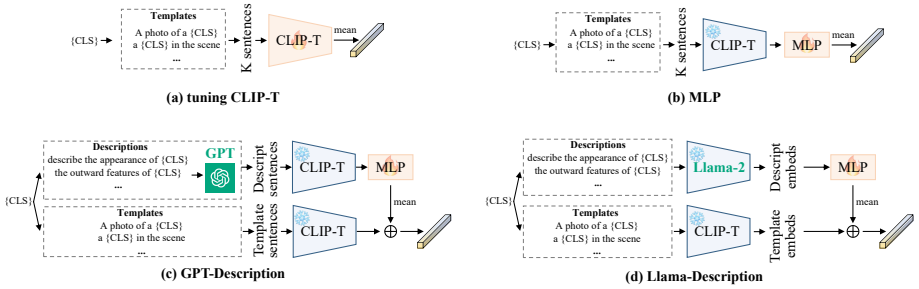

Fig. 4: Comparisons between CLIP-T tuning strategies.

Table 4: Ablation of diverse designs of CLIP-Text optimization. We remove the Representation Compensation strategy and freeze CLIP-V for analysis. Note that fine-tuning CLIP-T requires excessive GPU memory, and thus it is infeasible (denoted as N/A) for the setting in the 2^{nd} row.

	A-847	**A-150**	**PC-459**	**PC-59**	**PAS-20**
frozen CLIP (baseline)	11.2	26.6	12.7	42.4	89.5
+ fine-tune CLIP-T	N/A	N/A	N/A	N/A	N/A
+ MLP	4.1	20.2	11.2	51.4	89.4
+ GPT-Description	11.9	28.2	13.3	42.6	90.6
+ Llama-Description	9.6	26.1	11.5	40.8	90.9
+ Content-Dependent Transfer	13.3	32.4	17.2	55.2	94.7

- **a. fine-tuning CLIP-T** We explore fine-tuning CLIP-T parameters to optimize the CLIP text representation. The category name ({CLS}) is first augmented to sentences by some templates [18, 34–37] and fed into CLIP-T However, fine-tuning CLIP-T (2^{nd} results in Tab. 4) requires excessive GPU memory (more than 8 NVIDIA A100 GPUs), which is unaffordable in our experiments.
- **b. MLP** An MLP layer is added after CLIP-T, with the MLP learning to project text embedding to fit segmentation distributions. Within this design, CLIP-T is frozen, greatly reducing GPU memory consumption compared with fine-tuning CLIP-T. According to the 3^{rd} results in Tab. 4, the performance suffers a significant drop on ADE20K (11.2→4.1, 26.6→20.2), while increasing on PC-59 (42.4→51.4). This could be attributed to the MLP layer losing CLIP's zero-shot capability and its inability to perceive novel categories effectively.
- **c. GPT-Description** We assume that the detailed description of {CLS} contains additional valuable information, helping to optimize CLIP-T distribution. To explore this, we leverage GPT-3.5 [1] to generate description sentences of one {CLS}. *e.g.*, if the instruction provided to GPT is: [Instruct] ="Please describe the appearance of *cat*." GPT responds the description sentences of *cat*: [Response] = "[-a rounded head; -a short snout; -triangular ears ...]" Then we use a frozen CLIP-T to generate the corresponding text embeddings, followed by an MLP layer to project the embeddings. Within this design, the performance is slightly improved: +0.7 on A-847, +2.0 on A-150 (4^{th} results in Tab. 4).
- **d. Llama-Description** In view of Large Language Models (LLMs) powerful text representation capability, we explore to use of the open-source LLM, Llama-2 [30], to generate descriptive text embeddings. After obtaining Llama and CLIP-T embeddings, we average them and train an MLP layer to project the Llama embeddings into the CLIP-T embeddings space. Our experimental results demonstrate that this design does not benefit the performance (5^{th} results in Tab. 4). The mIoU drops from 11.2 to 9.6 on A-847, 12.7→11.5 on

PC459. This decrease may be due to the fact that the LLMs' feature space is not aligned with the CLIP-V's visual feature space.

– **Content-Dependent Transfer** We propose the Content-Dependent Transfer to enhance CLIP Text embeddings conditioned on the input images. Details can be found in Sec. 4.3. As shown in the last results in Tab. 4, the Content-Dependent Transfer improves the performance on all five datasets: 11.2→13.3, 26.6→32.4, 12.7→17.2, 42.4→55.2, and 89.5→94.7, respectively.

Analysis of why LLMs do not work? OVS focuses on data-limited settings, examines the model's ability to segment arbitrary text after seeing a few classes. Therefore, effective image-text alignment of prior models (*e.g.*, CLIP) is crucial. Despite LLMs' strong text processing capabilities, their potential is not fully realized with limited data, resulting in incomplete image-text alignment. Thus, simply adapting LLMs to OVS is unsuitable and may require further research. **Note:** The descriptions of all categories in the training set can be obtained through one single pre-processing step. Therefore, in **c. & d.**, the additional computational cost during training can be ignored. More details of the templates and the designs for GPT and Llama-2 are provided in the *Appendix*.

Table 5: Ablations of the Representation Compensation strategy. The Content-Dependent Transfer is removed. The best results are highlighted with red, and the default settings are highlighted with gray background.

	A-847	A-150	PC-59
None	14.6	34.8	57.1
Freeze {S0, 1}	14.6	34.7	57.0
Freeze {S0, 1, 2}	14.0	34.6	55.3
Freeze {S0, 1, 2, 3}	13.6	33.6	54.7

(a) Ablation of the **frozen stages** in CLIP-V.

	A-847	A-150	PC-59
Grid {1}	13.8	33.9	56.5
Grid {1, 2}	14.0	34.6	56.6
Grid {1, 2, 4}	14.6	34.8	57.1
Grid {1, 3, 6}	14.5	34.6	55.7

(b) Ablation of the AvgPooling **grid** in $\mathcal{L}_{rc}$.

Effect of Representation Compensation. We conduct ablation studies on Representation Compensation strategy in Fig. 5, here we remove the Content-Dependent Transfer for analysis.

– **Frozen stages in CLIP-V:** We explore the impact of fine-tuning units within CLIP-V. CLIP-V consists of 4 ConvNeXt stages {S0, S1, S2, S3}, which downsample the image features from $\frac{1}{4}$ to $\frac{1}{32}$. We start with fine-tuning the entire CLIP-V, and then freezing each stage sequentially, as detailed in Tab. 5a. Compared to fine-tuning the entire CLIP-V, freezing any stage causes performance degradation. Freezing S0-1, S0-2, S0-3 brings -0.1, -1.8, and -2.4 mIoU performance degradation respectively on PC-59, indicating that freezing S2 and S3 (depth convnext stages) has the most significant impact on the performance.

– **Effect of** AvgPooling **grids:** In Tab. 5b, we investigate how different multi-scale AvgPooling grids ($\{1\}$, $\{1, 2\}$, $\{1, 2, 4\}$, $\{1, 3, 6\}$) in $\mathcal{L}_{rc}$ impact per-

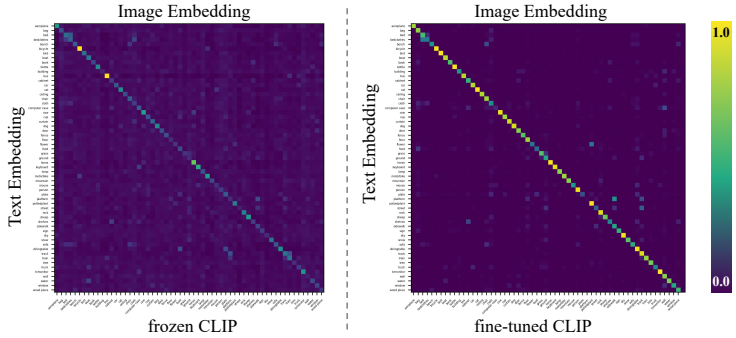

Fig. 5: Qualitative results. Normalized cosine similarity between the text embeddings and image embeddings of 59 classes in PC59. Text & image embeddings are generated by frozen CLIP (left) or our MAFT+ fine-tuned CLIP (right). The high similarity scores are highlighted in yellow, low similarity scores are shown in blue.

formance. Results show $\{1, 2, 4\}$ grids boost performance on A-150 to 34.8 mIoU, and achieve the best performance. Using $\{1, 3, 6\}$ grads results in -1.6 drops on PC-59, manifesting overly large AvgPooling grids compromises the model to learn region-level differences.

Table 6: Extending MAFT+ with ConvNeXt-Base CLIP. The best results are highlighted with red.

	A-847	**A-150**	**PC-459**	**PC-59**	**PAS-20**
FC-CLIP* [37]	12.7	31.1	12.5	54.3	93.8
MAFT +	13.2	33.6	14.2	55.9	93.9

Extending MAFT+ with ConvNeXt-Base CLIP. To showcase the efficacy and robustness of MAFT+, we conduct experiments using ConvNeXt-Base CLIP (Tab. 6). We include the results of FC-CLIP for comparison. MAFT+ outperforms FC-CLIP counterpart by a significant margin on all five datasets. This demonstrates that MAFT+ can easily transfer to other CLIP models.

5.4 Qualitative Study

Visualizations of similarity map. Fig. 5 presents the normalized similarity map between text and image embeddings. Including similarity map generated by frozen CLIP embeddings (left) and similarity map generated by fine-tuned CLIP embeddings (right). An observation can be obtained: The high similarity values of fine-tuned CLIP are mainly located on the diagonal of the similarity

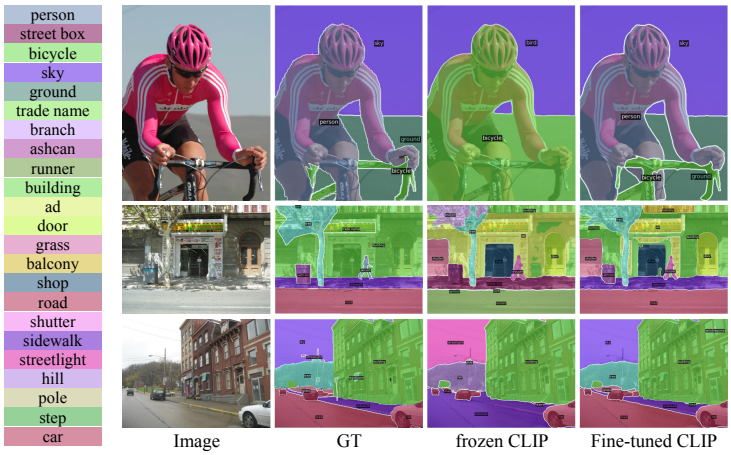

Fig. 6: Qualitative results. The results with the frozen CLIP and our MAFT+ fine-tuned CLIP are shown for comparasion.

map, indicating the collaborative optimization of CLIP-V and CLIP-T achieves better alignment of vision-text representation.

Qualitative analysis. We show some visual examples in Fig. 6. In some simple cases, the frozen CLIP results may contain background noise, and tend to classify multiple objects into one single class (*e.g.* the 1^{st} row, "bicycle"). The frozen CLIP is prone to misclassification when there are many categories in one image (the 3^{rd} row, "streetlight", "sidewalk", "hill"). Our fine-tuned CLIP collaboratively learns vision-text representation for segmentation tasks, which can significantly improve the segmentation results. In addition, the 2^{rd} row shows that our fine-tuned CLIP successfully segments "balcony", which is a reasonable outcome even though "balcony" does not appear in the ground-truth annotations. More visual samples are shown in the *Appendix*.

6 Conclusion

We rethink the issues in frozen CLIP paradigm and CLIP-V fine-tuning paradigm and propose a collaborative vision-text optimizing structure, MAFT+, for OVS. We introduce the Representation Compensation to review the original CLIP's representation to maintain the zero-shot capability of CLIP-V. And propose the Content-Dependent Transfer to optimize the text representation in a parameter-efficient way. Extensive experiments demonstrate our MAFT+ achieves superior performance on multiple open-vocabulary segmentation datasets.

Limitations. While MAFT+ optimizes the vision-test representation space of CLIP to fit the distribution of OVS, the optimization upper-bound is constrained by the capabilities of the pre-trained CLIP model. Addressing this limitation constitute our future research focus.

Acknowledgements

This work was supported in part by the National Key R & D Program of China (No. 2021ZD0112100), the National NSF of China (No.U23A20314).

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
2. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. Advances in Neural Information Processing Systems **32** (2019)
3. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1209–1218 (2018)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
5. Chen, X., Li, S., Lim, S.N., Torralba, A., Zhao, H.: Open-vocabulary panoptic segmentation with embedding modulation. arXiv preprint arXiv:2303.11324 (2023)
6. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
7. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34** (2021)
8. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11583–11592 (2022)
9. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015)
10. Fang, Y., Zhu, F., Cheng, B., Liu, L., Zhao, Y., Wei, Y.: Locating noise is halfway denoising for semi-supervised segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16612–16622 (2023)
11. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI. pp. 540–557. Springer (2022)
12. Gu, Z., Zhou, S., Niu, L., Zhao, Z., Zhang, L.: Context-aware feature generation for zero-shot semantic segmentation. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1921–1929 (2020)
13. Han, K., Liu, Y., Liew, J.H., Ding, H., Wei, Y., Liu, J., Wang, Y., Tang, Y., Yang, Y., Feng, J., et al.: Global knowledge calibration for fast open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
14. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 603–612 (2019)

15. Huang, Z., Wei, Y., Wang, X., Liu, W., Huang, T.S., Shi, H.: Alignseg: Feature-aligned segmentation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(1), 550–557 (2021)
16. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., et al.: Openclip, july 2021. If you use this software, please cite it as below **2**(4), 5 (2021)
17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
18. Jiao, S., Wei, Y., Wang, Y., Zhao, Y., Shi, H.: Learning mask-aware clip representations for zero-shot segmentation. Advances in Neural Information Processing Systems **36** (2023)
19. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9404–9413 (2019)
20. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
21. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=RriDjddCLN
22. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7070 (2023)
23. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)
24. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 891–898 (2014)
25. Qin, J., Wu, J., Yan, P., Li, M., Yuxi, R., Xiao, X., Wang, Y., Wang, R., Wen, S., Pan, X., et al.: Freeseg: Unified, universal and open-vocabulary image segmentation. arXiv preprint arXiv:2303.17225 (2023)
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
28. Shaban, A., Bansal, S., Liu, Z., Essa, I., Boots, B.: One-shot learning for semantic segmentation. arXiv preprint arXiv:1709.03410 (2017)
29. Sun, Y., Chen, Q., He, X., Wang, J., Feng, H., Han, J., Ding, E., Cheng, J., Li, Z., Wang, J.: Singular value fine-tuning: Few-shot segmentation requires few-parameters fine-tuning. arXiv preprint arXiv:2206.06122 (2022)
30. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

31. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8256–8265 (2019)
32. Xiao, J.W., Zhang, C.B., Feng, J., Liu, X., van de Weijer, J., Cheng, M.M.: Endpoints weight fusion for class incremental semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7204–7213 (2023)
33. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. CVPR (2022)
34. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2955–2966 (2023)
35. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2945–2954 (2023)
36. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIX. pp. 736–753. Springer (2022)
37. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. Advances in Neural Information Processing Systems **36** (2023)
38. Zhang, C.B., Xiao, J.W., Liu, X., Chen, Y.C., Cheng, M.M.: Representation compensation networks for continual semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7053–7064 (2022)
39. Zhang, G., Navasardyan, S., Chen, L., Zhao, Y., Wei, Y., Shi, H., et al.: Mask matching transformer for few-shot segmentation. Advances in Neural Information Processing Systems **35**, 823–836 (2022)
40. Zhang, G., Wang, L., Kang, G., Chen, L., Wei, Y.: Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19148–19158 (2023)
41. Zhang, Z., Gao, G., Fang, Z., Jiao, J., Wei, Y.: Mining unseen classes via regional objectness: A simple baseline for incremental segmentation. Advances in Neural Information Processing Systems **35**, 24340–24353 (2022)
42. Zhang, Z., Gao, G., Jiao, J., Liu, C.H., Wei, Y.: Coinseg: Contrast inter-and intra-class representations for incremental segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
43. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2881–2890 (2017)
44. Zheng Ding, Jieke Wang, Z.T.: Open-vocabulary universal image segmentation with maskclip. In: International Conference on Machine Learning (2023)
45. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 633–641 (2017)