

# Generalization Boosted Adapter for Open-Vocabulary Segmentation

Wenhai Xu, Changwei Wang, Xuxiang Feng, Rongtao Xu\*, Longzhao Huang, Zherui Zhang, Li Guo,  
Shibiao Xu\*, Member, IEEE

**Abstract**—Vision-language models (VLMs) have demonstrated remarkable open-vocabulary object recognition capabilities, motivating their adaptation for dense prediction tasks like segmentation. However, directly applying VLMs to such tasks remains challenging due to their lack of pixel-level granularity and the limited data available for fine-tuning, leading to overfitting and poor generalization. To address these limitations, we propose Generalization Boosted Adapter (GBA), a novel adapter strategy that enhances the generalization and robustness of VLMs for open-vocabulary segmentation. GBA comprises two core components: (1) a Style Diversification Adapter (SDA) that decouples features into amplitude and phase components, operating solely on the amplitude to enrich the feature space representation while preserving semantic consistency; and (2) a Correlation Constraint Adapter (CCA) that employs cross-attention to establish tighter semantic associations between text categories and target regions, suppressing irrelevant low-frequency “noise” information and avoiding erroneous associations. Through the synergistic effect of the shallow SDA and the deep CCA, GBA effectively alleviates overfitting issues and enhances the semantic relevance of feature representations. As a simple, efficient, and plug-and-play component, GBA can be flexibly integrated into various CLIP-based methods, demonstrating broad applicability and achieving state-of-the-art performance on multiple open-vocabulary segmentation benchmarks.

**Index Terms**—Open-vocabulary segmentation, CLIP, Adapter.

## I. INTRODUCTION

Image segmentation is one of the most classic and fundamental problems in computer vision [1]–[3] and autonomous driving [4]–[7], aiming to assign a semantic category to every pixel. Modern segmentation methods [8]–[10] rely heavily on large-scale annotated data, but typical datasets contain only

\*Rongtao Xu and Shibiao Xu are the corresponding authors (xurongtao2019@ia.ac.cn; shibiaoxu@bupt.edu.cn).

Wenhai Xu, Shibiao Xu, Longzhao Huang, Zherui Zhang and Li Guo are with School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. Changwei Wang is with the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, 250013, China; Shandong Provincial Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China. Xuxiang Feng is with the Aerospace Information Research Institute, Chinese Academy of Sciences, China. Rongtao Xu is with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

This work is supported by Beijing Natural Science Foundation No. JQ23014, and by the Science and Disruptive Technology Program, AIRCAS (No. E2Z218020F), and by the National Natural Science Foundation of China (Nos. 62271074, 62071157, 62302052, 62171321 and 62162044).

tens to hundreds of categories. The high cost of data collection and annotation limits the application of these methods in practical scenarios that require handling open-vocabulary objects.

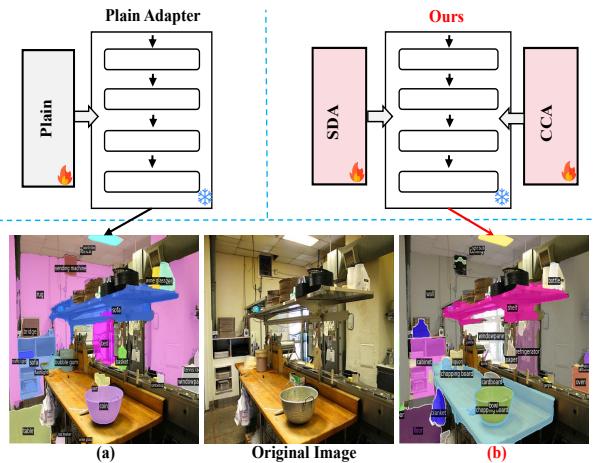


Fig. 1. We use FC-CLIP as the baseline to investigate the performance of different adapter methods in dense prediction tasks. As shown in (a), the model with plain adapter incorrectly identifies the object as a “sofa” and further misclassifies the “wall” as a “rug” due to false associations. In contrast, our proposed GBA method accurately recognizes the ‘shelf’ and ‘wall’ in the image, as illustrated in (b). This improvement can be attributed to the GBA’s ability to enhance the model’s generalization capability and suppress false associations for dense prediction tasks.

Recently, vision-language models (VLMs) [11], [12] have gained significant attention due to their remarkable open-vocabulary object recognition capability. This tremendous success has motivated us to explore their adaptability to the segmentation task. VLMs have demonstrated outstanding feature representation capabilities through cross-modal contrastive learning. However, such representation lacks pixel-level granularity, making it challenging to directly apply to dense prediction tasks. Inspired by work in natural language processing [13], [14], a series of methods have been proposed in the vision domain, aiming to efficiently adapt VLMs to open-vocabulary segmentation tasks. Existing methods include prompt learning [15], [16] and refining learned representations using adapters [17], [18]. Despite the progress made by prompt tuning and adapter techniques in open-vocabulary tasks, fine-tuning approaches that introduce a small number of learnable parameters for downstream tasks still have limitations due to the significantly smaller size of the fine-tuning dataset compared to the pre-training dataset used for VLMs: (1) Fine-tuning models with limited data may lead to overfitting on

specific patterns in the training samples, making it difficult to effectively learn general visual concepts, resulting in overfitting or poor generalization performance [19], [20]. (2) Training data contains a large amount of “noise” information irrelevant to semantic categories, such as background textures and object styles [21], [22]. Direct fine-tuning may cause the model to overly focus on this irrelevant information and establish “false associations” with the correct categories, undermining the original robustness and generalization capabilities of VLMs.

These methods are limited in their performance on open-vocabulary segmentation tasks because they cannot significantly alter the fundamental visual representations encoded in the model (e.g., CLIP), often leading to overfitting to the training data or poor out-of-distribution generalization. As illustrated in Fig.1 (a), the model with a plain adapter incorrectly recognizes a shelf, consequently misidentifying other objects as well. In this paper, we propose introducing adapter modules on top of the CLIP backbone network to enhance CLIP’s generalization and robustness in open-vocabulary segmentation tasks, as shown in Fig.1 (b). To achieve this objective, we devise a novel adapter strategy, termed **Generalization Boosted Adapter** (GBA), which comprises two core components: First, we introduce a **Style Diversification Adapter (SDA)** that decouples features into amplitude and phase components through Fourier transform, corresponding to style and content information, respectively. We operate solely on the amplitude component, aiming to alter the style while preserving semantic consistency, thereby enriching the feature space representation of the visual encoder. Second, we incorporate a **Correlation Constraint Adapter (CCA)** into the deeper layers of the visual encoder. This adapter employs a cross-attention mechanism to establish a tighter semantic association between text categories and target regions in the image, guiding the suppression of irrelevant low-frequency “noise” information. Consequently, it avoids erroneous associations between correct categories and irrelevant information, thus enhancing category matching accuracy. Moreover, GBA serves as a simple and efficient modular component that can be flexibly integrated into various CLIP-based methods, rendering it a universal solution for addressing diverse open-vocabulary downstream tasks. Through the synergistic effect of the shallow SDA and the deep CCA, the proposed GBA method effectively alleviates overfitting issues and enhances the semantic relevance of feature representations, thereby significantly boosting the performance of open-vocabulary segmentation tasks. As a simple, efficient, and plug-and-play component, GBA demonstrates broad applicability and can be readily employed in various CLIP-based open-vocabulary downstream tasks. Our paper makes the following contributions:

(i) To mitigate the overfitting issue caused by limited fine-tuning data, we propose a **Style Diversification Adapter** (SDA) to enhance feature diversity, preventing the model from overly memorizing specific patterns in the training data and improving generalization, while maintaining content invariance through a content consistency loss.

(ii) To address the issue of visual feature representation quality, We propose a **Correlation Constraint Adapter** (CCA) that implicitly enforces a false association constraint

by suppressing irrelevant low-frequency information, thereby enhancing the semantic relevance between text categories and downstream images, which greatly benefits the category matching process.

(iii) The proposed GBA method leverages the synergistic effect of two different adapter strategies in a plug-and-play manner, achieving state-of-the-art results on multiple benchmarks.

## II. RELATED WORKS

### A. Vision Language Models

Vision-language models aim to learn generic representations that align visual and textual information. Early works, such as UNITER [23] and Oscar [24], employed a two-stage approach: first, they pretrained object detectors on smaller datasets to extract semantic representations, and then fine-tuned these representations on downstream tasks like visual question answering (VQA) and image captioning. However, these methods required carefully designed fusion encoders to integrate cross-modal interactions. Inspired by the success of pretraining in computer vision (CV) and natural language processing (NLP), many researchers have proposed approaches to pretrain large-scale models that process both visual and language modalities simultaneously. A typical visual-language model consists of four key components: a visual encoder, a language encoder, a fusion encoder, and loss functions. Building on the success of foundation models in CV and NLP domains, such as BERT [25] and Vision Transformers [26], [27], the multimodal learning community has leveraged these large-scale foundation models to boost performance [28], [29]. For example, VisualBERT [30], OSCAR [24], and Uniter [23] utilize BERT [25] to preprocess raw texts and have achieved impressive results on multimodal tasks like VQA. Recent breakthroughs in large-scale visual-language models, such as CLIP [12] and ALIGN [11], have demonstrated that dual-encoder models pretrained on large-scale image-text pair datasets can learn representations with cross-modal alignment. These models have rapidly advanced zero-shot and open-vocabulary downstream tasks. [12] and [11] have shown that visual-language contrastive learning can generate transferable features for downstream tasks, and the multimodal interaction can be well explained by simply computing the dot product between visual and language embeddings. Recently, some studies have applied large-scale visual-language models to open-vocabulary segmentation and other downstream tasks, achieving impressive results and confirming the benefits of pretraining on large-scale text-image pairs. Inspired by these methods, the current work explores the introduction of specially designed dual adapters, allowing models to produce superior open-vocabulary classification segmentation results while preserving the cross-modal alignment capabilities of large-scale vision-language models.

### B. Adapting Vision-Language Models

The concept of adapters, initially introduced in the field of natural language processing (NLP) for fine-tuning large pre-trained models on downstream tasks, has recently garnered significant attention in the vision-language domain. Large-scale

pre-trained language models have demonstrated remarkable success in various natural language tasks, such as question answering, sentence completion, and language translation [31], [32]. However, due to the immense scale of these models, training or fine-tuning them is prohibitively expensive. To address this issue, researchers have proposed several efficient adaptation methods, including designing lightweight adapter modules, bias adjustment calibration, or learning task-specific soft text prompts, enabling scholars to leverage the power and generality of large pre-trained language models on downstream tasks of interest. In the vision-language domain, methods like CLIP [12] learn the embedded representations of images and text through contrastive learning, where the embeddings are encoded by independent image and language encoders. Essentially, CLIP [12] learns the representations of both modalities by pulling the image representations closer to their paired text representations while pushing away the text embeddings corresponding to different images. The advantage of these methods lies in their ability to learn from vast amounts of unstructured image and text data available on the internet, allowing vision-language models to be applied to various downstream tasks in zero-shot or few-shot settings. Inspired by the adaptation of natural language models, two primary approaches have been proposed to adapt CLIP-like models to downstream vision tasks: prompt learning and feature adaptation. Prompt learning methods, such as CoOp [19] and CoCoOp [33], aim to automatically construct text prompts by optimizing learnable vectors. In contrast, feature adaptation methods, such as CLIP-Adapter [34] and Tip-Adapter [17], directly adjust the representations extracted from the visual and text encoders of CLIP-like models. CLIP-Adapter [34] adds a lightweight fully-connected neural network adapter applied to frozen CLIP [12] features and fine-tunes the parameters on the downstream task of interest with limited supervision. Tip-Adapter achieves better results by constructing a key-value cache model from few-shot examples and performing fewer fine-tuning steps. Furthermore, a tune-free version of Tip-Adapter [17] has been proposed, which adapts faster during training but yields inferior performance. Although these adapter methods are relatively efficient, they are insufficient for making large-scale changes to the fundamental representations extracted from the backbone visual or text encoders, which may pose a problem when the evaluation task differs significantly from the distribution encountered during training. Recently, several works have attempted to add adapters to visual foundation models to better perform pixel-level prediction tasks. SAM-Adapter [35] adapts SAM to specific scenarios, such as shadow detection and camouflaged object detection, by adding learnable adapters. SAN [36] improves the performance of open-vocabulary semantic segmentation by adding two parallel side networks as adapters on top of a frozen backbone network. Despite the proven benefits of the aforementioned basic adapters in adapting pre-trained models to downstream tasks, they still suffer from overfitting issues due to the relatively small size of the downstream datasets used for fine-tuning. Based on this premise, we have specifically designed novel adapter strategies to address the problems of feature robustness and spurious correlations that arise when adapting to downstream

segmentation tasks, stemming from the inherent limitations of CLIP [12] and the small scale of downstream datasets.

### C. Open Vocabulary Segmentation

Deep learning [37]–[41] and image segmentation have recently achieved remarkable success [42]–[47]. Open vocabulary segmentation aims to segment target categories that are unseen during training, which remains a challenging task due to the limited availability of annotated data and the vast diversity of objects in the real world. The existing approaches can be divided into two aspects: mapping visual features into semantic space and cross-modal alignment with pre-trained models [48], [49]. For the mapping aspect, SPNet [50] encodes visual features to the semantic embedding space and then projects each pixel feature to predict probabilistic outcomes through a fixed semantic word encoding matrix. ZS3Net [51] generates the pixel-level features of unseen classes in the semantic embedding space and adopts the generated features to supervise a visual segmentation model. STRICT [52] introduces a self-training technique into SPNet to improve the segmentation performance of unseen classes. Recent advancements in large-scale visual language modeling have significantly contributed to the progress of open-vocabulary semantic segmentation. LSeg [53] learns a CNN model to compute per-pixel image features to match with the text embeddings embedded by the pre-trained text model. ZegFormer [48] and ZSSeg [49] leverage the visual model to generate the class-agnostic masks, and use the pre-trained text encoder to retrieve the unseen class masks. XPM [54] utilizes the region-level features to match CLIP-based text embeddings to accomplish the open vocabulary instance segmentation. Some of these studies [55], [56] fine-tuned the visual language pre-training model, but would compromise the cross-modal alignment ability of the visual language model. However, fine-tuning the visual language pre-training model, as done in some studies [55], [56], may compromise the cross-modal alignment ability of the model.

To address this issue, various approaches have been proposed. SimSeg [57] and MaskCLIP [58] introduce a two-phase framework that first generates class-independent masks and then assigns classes to the masks using a frozen CLIP. ODISE [59] and FreeSeg [60] enhance the quality of open-vocabulary semantic segmentation by incorporating a diffusion model and a multi-granularity concepts encoder, respectively. FC-CLIP [61] utilizes a shared frozen convolutional CLIP backbone within an efficient single-stage framework. SAN [36] extends the frozen CLIP with two side branches: one for predicting mask proposals and another for predicting attentional bias, which is used to identify mask classes. Building upon these advancements, our work further investigates the adaptability of the frozen-parameter CLIP model to the open-vocabulary semantic segmentation task. We propose leveraging a novel dual adapter strategy to enrich the feature representation space and alleviate spurious associations between categories and noise, thereby enhancing the generalization performance of open-vocabulary semantic segmentation.

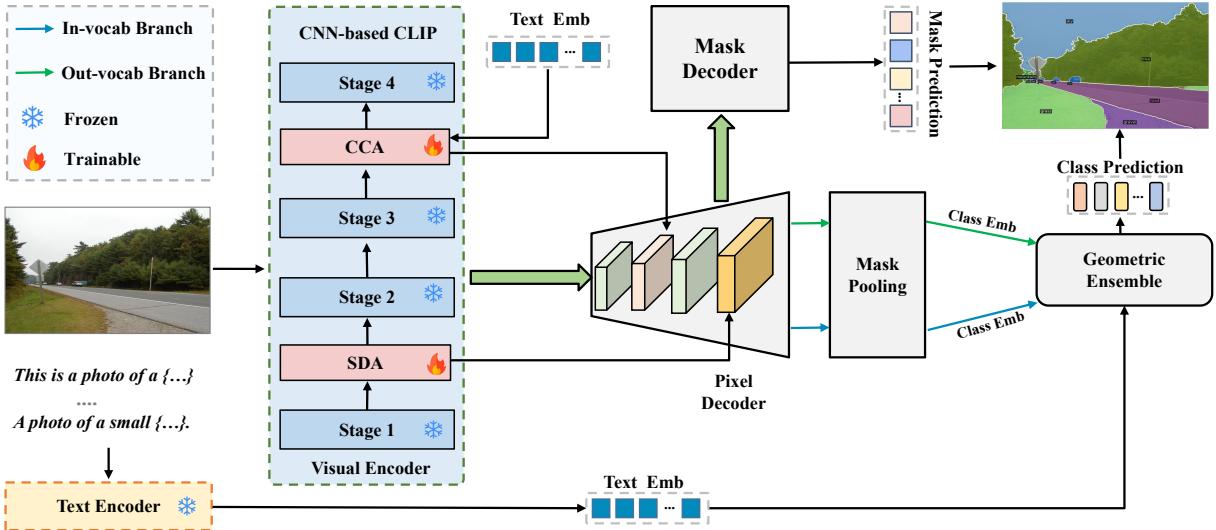


Fig. 2. **Overview of the GBA framework.** Single-stage open-vocabulary segmentation methods akin to FC-CLIP comprise three key components: a mask generator, an in-vocabulary classifier, and an out-of-vocabulary classifier. All these components are constructed based on features extracted from a frozen CNN-based CLIP backbone that employs the proposed two learning feature augmentation adapters, namely, the Style Diversification Adapter (SDA) and the Correlation Constraint Adapter (CCA).

### III. METHOD

#### A. Task Definition

Given an image  $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ , where  $H$  and  $W$  denote the height and width of the image respectively, open-vocabulary semantic segmentation aims to segment the image into a set of masks with associated semantic classes:

$$\mathcal{Y} = (\mathcal{M}_i, \mathcal{C}_i)_{i=1}^N. \quad (1)$$

where  $\mathcal{M}_i \in \{0, 1\}^{H \times W}$  represents the ground truth mask and  $\mathcal{C}_i \in \mathcal{S}_{train} \cup \mathcal{S}_{test}$  denotes the corresponding ground truth class label.

Open-vocabulary semantic segmentation is more challenging than traditional image segmentation tasks [62], [63] because the inference classes are not observed during training. During evaluation, the test categories  $\mathcal{C}_{test}$  are different from  $\mathcal{C}_{train}$ , containing novel categories not seen in training, i.e.,  $\mathcal{C}_{train} \neq \mathcal{C}_{test}$ .

#### B. Baseline.

We adopt the FC-CLIP [61] as our baseline and integrate our proposed adapters as lightweight components. The image encoder is configured as a ConvNext-Large model, comprising four stages. Each stage contains a different number of blocks: 3 in the first, 3 in the second, 27 in the third, and 3 in the fourth. In contrast, the text encoder is structured as a 16-layer transformer, each layer being 768 units wide and featuring 12 attention heads. We harness the power of multi-scale features extracted by the image encoder. These features are represented as feature maps of varying widths and scales: a 192-wide feature map downscaled by a factor of 4, a 384-wide map downscaled by 8, a 768-wide map downscaled by 16, and a 1536-wide map downscaled by 32. FC-CLIP incorporates a frozen-parameter convolutional CLIP backbone (ConvNeXt [64]) into the Mask2Former [8] segmentation

pipeline, achieving a simple yet effective single-stage open-vocabulary semantic segmentation approach.

#### C. Generalization Boosted Adapter: Diversifying Features and Correlation Constraint

CLIP is pre-trained through image-level contrastive learning, and thus using the frozen CLIP visual encoder to extract features for downstream dense prediction tasks leads to suboptimal results. This is because CLIP utilizes image-level supervision signals during pre-training, failing to capture pixel-level detail features. Moreover, when adapting CLIP features to dense prediction tasks, their spatial attention is commonly scattered across different regions of the image, resulting in spurious correlations when fine-tuning the learnable parameters on limited data. To alleviate the above issues, a common approach is to insert learnable adapters into CLIP. Traditional adapter techniques primarily consist of linear layers, downsampling layers, nonlinear activation layers, upsampling layers, and skip connections. However, due to the small number of parameters in these standard adapters, they are prone to overfitting during training, excelling at handling seen classes but lacking generalization capability for unseen open-vocabulary classes. Based on these observations, we propose the Generalization Boosted Adapter (GBA) dual-adapter design paradigm, as shown in Fig. 2. GBA comprises two adapter modules with distinct functionalities, respectively enhancing spatial feature style diversity adaptability and suppressing spurious correlations, aiming to improve the model's generalization ability for open-vocabulary semantic segmentation.

In our proposed lightweight plain adapter architecture, we introduce multiple convolutional layers with varying receptive fields to enhance the feature extraction capability of the adapter. Specifically, the feature maps are sequentially processed by three convolutional layers with kernel sizes of  $3 \times 3$ ,  $5 \times 5$ , and  $7 \times 7$ , respectively. We compute the average of these

three layers and employ a  $1 \times 1$  convolution to aggregate the features. After applying a SiLU layer for non-linear activation, the feature maps undergo a feature augmentation strategy module. Additionally, we incorporate residual connections. Finally, a simple projection layer yields the output feature maps.

As illustrated in Fig. 3, our proposed two adapter modules adopt a similar base architecture but employ different feature augmentation strategies. In the following sections, we will elaborate on the design details and working mechanisms of these two strategies.

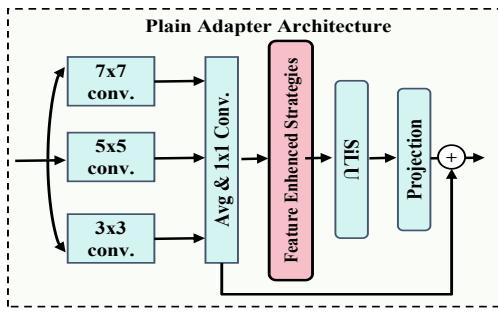


Fig. 3. **The detail of plain adapter.** Adapters exhibiting diverse generalization capacities can be realized through the utilization of varied feature enhanced strategies.

**1) Style Diversification Adapter:** Extending feature-level style variations has been recognized as an effective approach to enhance model robustness. Numerous studies, such as mixstyle and AdaIN, have been inspired by the observation that the statistical feature distributions in convolutional neural networks (CNNs) contain crucial style information. These methods synthesize stylized features by adjusting feature statistics, thereby increasing feature diversity. Well-known normalization techniques, including normalization (BN) [65], layer normalization (LN) [66], and instance normalization (IN) [67], are commonly employed in these approaches since normalization statistics encapsulate style information, and normalization can effectively extract style from features. However, when altering or removing style, semantic content may also undergo changes [68], [69]. Recent research [70], [71] has discovered that utilizing Fourier transform can effectively decompose input images into amplitude and phase components, which represent style and content, respectively. In tackling the single-domain generalization challenge, Wang [72] leverages feature frequency modulation to synthesize images with diverse styles, effectively bolstering the generation of hard samples. Based on the aforementioned insights, we propose a novel frequency-domain decomposition-based feature diversification strategy. This strategy aims to enrich style diversity while simultaneously preserving semantic integrity. By mapping features to the frequency domain and manipulating amplitude and phase components independently, we can adjust style and content separately, thereby generating rich style variations while maintaining semantic content consistency, as shown in Fig.4 (a).

First, we extract the base style features of the input  $\mathbf{x}$  and perform frequency decomposition. We compute the mean

$\mu_{base}$  and standard deviation  $\sigma_{base}$  of the base style features of  $\mathbf{x}$ , where  $C$  denotes the number of channels, and  $H$  and  $W$  denote the height and width, respectively. Simultaneously, we apply the Fast Fourier Transform (FFT) to the original sample  $x \in \mathbb{R}^{C \times H \times W}$  to obtain the frequency domain features  $F(x)$ , which we then decompose into the amplitude  $\mathbf{a}$  and phase  $\mathbf{p}$ :

$$\mathbf{a} = \sqrt{F(x)_{\text{real}}^2 + F(x)_{\text{img}}^2}, \quad (2)$$

$$\mathbf{p} = \arctan \left( \frac{F(x)_{\text{img}}}{F(x)_{\text{real}}} \right), \quad (3)$$

Here,  $F_{\text{real}}$  and  $F_{\text{img}}$  represent the real and imaginary parts of the Fourier coefficients, respectively. Next, we perform normalization and style fusion.

We then element-wise multiply  $\mu_{base}$  and  $\sigma_{base}$  with the combination weights  $\mathbf{W} \in \mathbb{R}^C$  sampled from the Dirichlet distribution  $B(\alpha_1, \dots, \alpha_C)$  to generate the amplitude features of the new style:

$$\mu = \mathbf{W} \cdot \mu_{base}, \quad \sigma = \mathbf{W} \cdot \sigma_{base}, \quad (4)$$

$$\mathbf{a}_{\text{new}} = \sigma \cdot \mathbf{a} + \mu, \quad (5)$$

Finally, we recombine the new style amplitude  $\mathbf{a}_{\text{new}}$  with the phase  $\mathbf{p}$  of the original sample and perform the Inverse Fast Fourier Transform (IFFT) to generate the sample  $\tilde{\mathbf{x}} \in \mathbb{R}^{C \times H \times W}$  with the new style:

$$\tilde{\mathbf{x}} = \text{IFFT}(\text{Compose}(\mathbf{a}_{\text{new}}, \mathbf{p})). \quad (6)$$

Throughout this process, by keeping the phase of the original sample unchanged, we ensure content consistency and achieve the goal of preserving content information during style transformation. Meanwhile, by normalizing the amplitude features and fusing different styles, we generate samples with new styles.

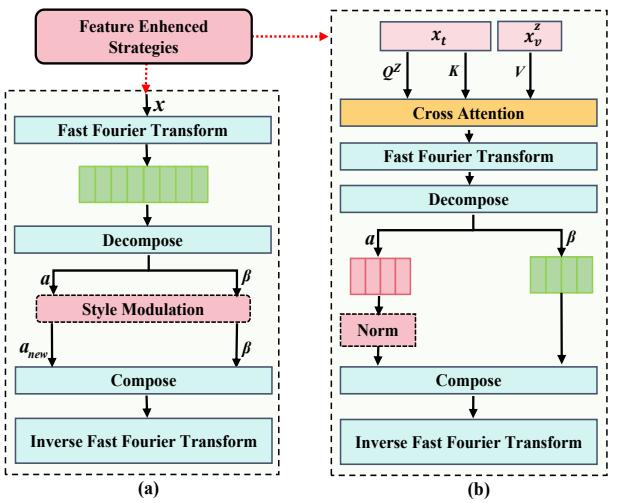


Fig. 4. The details of Style Diversification Adapter (SDA) (a) and Correlation Constraint Adapter (CCA) (b).

**2) Correlation Constraint Adapter:** False correlation, often caused by background elements irrelevant to the target label, is a common issue in visual recognition tasks. Previous research

[73], [74] has revealed that high-frequency components, such as object edges and contours, exhibit stronger associations with semantic features compared to the relatively lower frequencies of backgrounds and object surfaces [75]. Inspired by this insight, we propose a novel approach to mitigate the impact of false correlation on model performance, as shown in Fig.4 (b). By emphasizing learning from high-frequency components and mitigating the risk of learning erroneous correlations, the model can more effectively capture the semantic features of objects. To achieve this goal, we introduce a cross-attention mechanism for semantic interaction to model the relevance between text embeddings and multi-scale visual features, thereby guiding the model to learn semantically-relevant high-frequency information. Specifically, the semantic interaction can be formulated as:

$$Attn(\mathbf{Q}^z, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}^z \mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V}, \quad (7)$$

$$\mathbf{Q}^z = \phi_q(\mathbf{X}_v^z), \quad \mathbf{K} = \phi_k(\mathbf{X}_t), \quad \mathbf{V} = \phi_v(\mathbf{X}_t), \quad (8)$$

where  $\mathbf{X}_v^z$  denotes the visual features from the  $z$ -th layer of the decoder in the masked extractor, and  $\mathbf{X}_t$  represents the text category features.  $\mathbf{Q}^z, \mathbf{K}, \mathbf{V}$  represent the query, key, and value embeddings generated by the projection layers  $\phi_q, \phi_k, \phi_v$ , respectively.  $\sqrt{d_k}$  is the scaling factor. The attention relation is then utilized to enhance the visual features:

$$\hat{\mathbf{X}}_v^z = \mathcal{L}\{Attn[\phi_q(\mathbf{X}_v^z), \phi_k(\mathbf{X}_t), \phi_v(\mathbf{X}_t)]\}, \quad (9)$$

where  $\mathcal{L}$  represents the output projection layer. The enhanced visual features  $\hat{\mathbf{X}}_v^z$  facilitate emphasizing the visual information relevant to the given text category, thereby better capturing the high-frequency semantic features of objects.

Subsequently, we perform operations in the Fourier frequency domain to facilitate the model's utilization of high-frequency information. By operating in the frequency domain, we can exert finer-grained control over the model's attention to different frequency components, thereby more effectively mitigating the impact of error correlation. Specifically, we first apply the Fast Fourier Transform (FFT) to the visual features  $\hat{\mathbf{X}}_v^z \in \mathbb{R}^{C \times H \times W}$ , mapping them into the frequency domain to obtain the frequency domain features  $F(\hat{\mathbf{X}}_v^z)$ . We then decompose  $F(\hat{\mathbf{X}}_v^z)$  into the amplitude component  $\mathbf{a}$  and the phase component  $\mathbf{p}$ . The phase component  $\mathbf{p}$  primarily encodes high-frequency details such as edges and textures in the image, while the amplitude component  $\mathbf{a}$  predominantly reflects low-frequency global characteristics such as smoothness and contrast. To enhance the influence of high-frequency structural information, we perform normalization on the amplitude component  $a$ :

$$\mathbf{a}_{\text{norm}} = \frac{\mathbf{a} - \mu(\mathbf{a})}{\sigma(\mathbf{a})}, \quad (10)$$

where  $\mu(\mathbf{a})$  and  $\sigma(\mathbf{a})$  represent the mean and standard deviation of  $a$ , respectively. Through this normalization operation, we can implicitly amplify the relative importance of high-frequency structural information while suppressing the impact of low-frequency global features. Finally, we recombine the

normalized amplitude  $\mathbf{a}_{\text{norm}}$  with the phase  $\mathbf{p}$  of the original visual features and perform the Inverse Fast Fourier Transform (IFFT) to generate the processed visual features  $\hat{\mathbf{X}} \in \mathbb{R}^{C \times H \times W}$ :

$$\hat{\mathbf{X}} = \text{IFFT}(\text{Compose}(\mathbf{a}_{\text{norm}}, \mathbf{p})). \quad (11)$$

By combining the cross-attention mechanism and frequency domain analysis, our approach effectively mitigates the impact of false correlation on model performance while enhancing the model's ability to capture and comprehend visual and textual information directly relevant to the task.

## IV. EXPERIMENT

### A. Implementation Details

**Network Architecture.** Our network architecture is based on the FC-CLIP [61] baseline, employing a ConvNeXt-Large CLIP backbone [64] pretrained on the LAION-2B dataset [76] using OpenCLIP. The pixel and mask decoders follow the default settings of Mask2Former [77] for generating category-independent masks. For in-vocabulary classification, we pool pixel features from the final decoder output based on mask predictions, as described in [61], [77], and compute the final classification logits via a matrix multiplication between the predicted class embeddings and the corresponding text embeddings of each class name.

**Training and Optimization.** For the training process, we follow the approach of [8] and adopt the same training recipe and losses without any special design. The optimization is performed using the AdamW optimizer [78] with a weight decay of 0.05. The input images are cropped to a size of  $1024 \times 1024$ , and the initial learning rate is set to  $1 \times 10^{-4}$  with a multi-step decay schedule employed to dynamically adjust the learning rate. The training batch size is 16, and the model is trained for 50 epochs on the COCO panoptic training set [79].

**Inference Procedure.** At inference time, input images are resized such that the shorter side is 800 pixels, while ensuring the longer side does not exceed 1333 pixels. Mask predictions are merged using a mask-wise merging scheme [8]. The out-vocabulary classifier operates on the frozen CLIP backbone features, and the final classification results are obtained through the geometric ensembling of in- and out-vocabulary classifiers [59], [61]. We also incorporate prompt engineering methods from [59] and prompt templates from [55] using their default settings.

**Evaluation Metrics and Datasets.** We primarily evaluate our model on the open-vocabulary semantic segmentation and open-vocabulary panoptic segmentation tasks. For open-vocabulary semantic segmentation, we perform zero-shot evaluation on the COCO [79], ADE20K [80], and PASCAL [81] datasets. The open-vocabulary semantic segmentation results are evaluated using the mean Intersection-over-Union (mIoU) metric. For open-vocabulary panoptic segmentation, we evaluate the model on the COCO [79] and ADE20K [80] datasets. We report the panoptic quality (PQ), semantic quality (SQ), and recognition quality (RQ) for open-vocabulary panoptic segmentation.

TABLE I  
**OPEN-VOCABULARY PANOPTIC SEGMENTATION PERFORMANCE.** “COCO P.” DENOTES THE COCO PANOPTIC DATASETS. “COCO” DENOTES THE COCO IMAGE DATASET. “IN 1K” DENOTES THE IMAGENET-1K IMAGE DATASET. WE REPORT PQ, SQ AND RQ FOR ALL DATASETS.

<b>Method</b>	<b>Training Data</b>	<b>COCO</b>			<b>ADE20K</b>		
		<b>PQ</b>	<b>SQ</b>	<b>RQ</b>	<b>PQ</b>	<b>SQ</b>	<b>RQ</b>
CutLER+STEGO [82]	IN 1K + COCO	12.4	64.9	15.5	-	-	-
U2Seg [83]	IN 1K + COCO	16.1	71.1	19.9	-	-	-
MaskCLIP [58]	COCO P.	30.9	-	-	15.1	70.5	19.2
ODISE [59]	COCO P.	55.4	-	-	22.6	-	-
FC-CLIP [61]	COCO P.	54.4	83.0	64.8	26.8	71.6	32.3
<b>Ours</b>	COCO P.	<b>57.5</b>	<b>83.7</b>	<b>67.9</b>	<b>29.6</b>	<b>74.0</b>	<b>35.8</b>

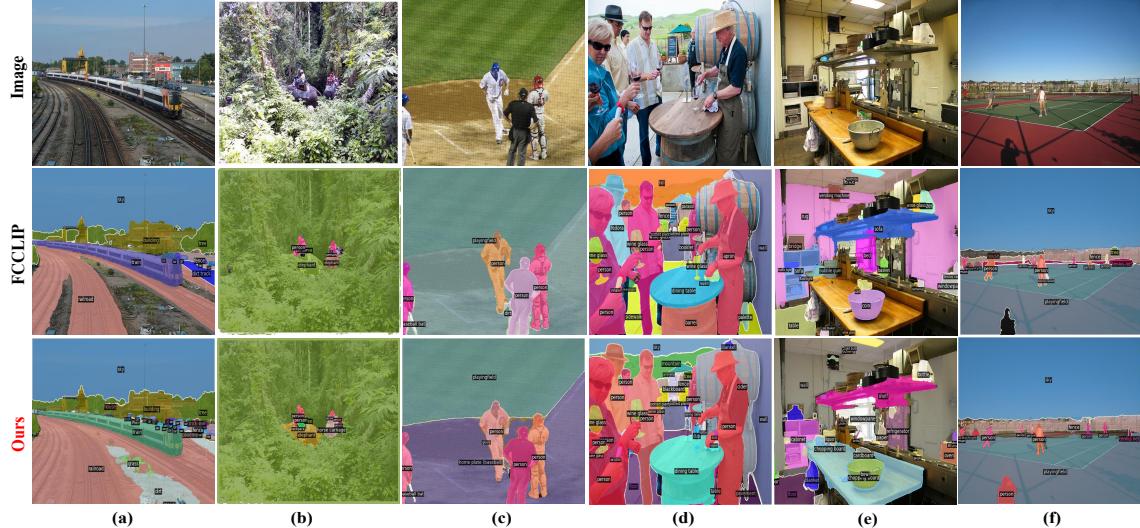


Fig. 5. **Qualitative Visualization of Open-Vocabulary Panoptic Segmentation.** To showcase the open-vocabulary recognition capability, we amalgamated class names from all datasets totaling approximately all classes, and conducted open-vocabulary inference directly.

## B. Main Results

1) *Open-vocabulary panoptic segmentation:* For open-vocabulary panoptic segmentation, we evaluate our method on the COCO [79] and ADE20K [80] datasets, as shown in Table I. Compared to other methods, our approach demonstrates superior performance on both datasets. On the COCO dataset, we achieve a PQ of 57.5%, surpassing the previous best method FC-CLIP [61] by 3.1%. Furthermore, our method yields an SQ of 83.7% and an RQ of 67.9%, outperforming FC-CLIP by 0.7% and 3.1%, respectively. These improvements indicate that our approach can effectively capture both the quality and completeness of panoptic segmentation masks. On the more challenging ADE20K dataset, our method also sets a new state-of-the-art, achieving a PQ of 29.6%, an SQ of 74.0%, and an RQ of 35.8%. Compared to the previous best method FC-CLIP, we obtain significant improvements of 2.8%, 2.4%, and 3.5% in PQ, SQ, and RQ, respectively. These results demonstrate the robustness and generalization capability of our approach in handling diverse semantic categories and complex scenes. In Figure 5, we visualize the qualitative results of the unified open-vocabulary panoptic segmentation. It can be observed that our method not only recognizes more instances but also discovers more categories, such as ‘grass’ and ‘dirty’ in (a), and ‘glove’ in (c). Furthermore, our method enhances the accuracy of object classification, as exemplified in (e).

2) *Open-vocabulary semantic segmentation:* As shown in Table II, we provide a comprehensive comparison of our GBA against prior works on a suite of benchmark datasets. These datasets include ADE20K [80] (including 150 and 847 class variants), PASCAL Context [93] (459 and 59 class variants), and PASCAL VOC [81] (with 20 and 21 classes). GBA demonstrates consistent performance improvements across all evaluated datasets. Specifically, on the more challenging ADE20K-847 dataset, GBA achieves 15.1% mIoU, outperforming the previous state-of-the-art method FC-CLIP [61] by 0.3 %. Similarly, on the PASCAL Context-459 dataset, GBA surpasses FC-CLIP by 0.3%, reaching 18.5% mIoU. These results highlight GBA’s exceptional ability in classifying diverse semantic categories. Furthermore, on the PASCAL VOC [81] benchmark datasets, GBA exhibits substantial improvements, achieving 95.8% and 84.5% mIoU on the 20 and 21 class variants, respectively. This indicates that our GBA can accurately capture class distinctions and fine-grained spatial structures. In Figure 6, we visualize the qualitative results of the unified open-vocabulary semantic segmentation. It can be observed from (b) that FC-CLIP [61] predicts ‘lake’ and ‘water’ separately, while our method GBA accurately segments the entire ‘lake’ region, demonstrating our method’s superior ability to recognize and segment complete regions. As shown in (a), (c), and (d), our method also discovers more categories. Moreover, our method exhibits higher accuracy in

TABLE II

**OPEN-VOCABULARY SEMANTIC SEGMENTATION PERFORMANCE.** WE MAINLY COMPARE WITH THE FULLY-SUPERVISED AND WEAKLY-SUPERVISED METHODS. “COCO S.”, “COCO P.” AND “COCO C.” DENOTE THE COCO STUFF, PANOPTIC AND CAPTION DATASETS. “O365” DENOTES THE OBJECT 365 DATASET. “M. 41M” DENOTES THE MERGED 41M IMAGE DATASET. WE REPORT MIoU FOR ALL DATASETS.

Method	Training Data	A-847	PC-459	A-150	PC-59	PAS-21	PAS-20
		mIoU (%)					
GroupViT [84] <i>CVPR'22</i>	GCC + YFCC	4.3	4.9	10.4	23.4	52.3	79.7
TCL [85] <i>CVPR'23</i>	GCC	-	-	14.9	30.3	51.2	77.5
OVSeg [86] <i>CVPR'23</i>	CC4M	-	-	5.6	-	53.8	-
SegCLIP [87] <i>ICML'23</i>	CC3M + COCO C.	-	-	8.7	-	52.6	-
CLIPPy [88] <i>ArXiv'23</i>	HQITP-134M	-	-	13.5	-	52.2	-
MixReorg [89] <i>ICCV'23</i>	CC12M	-	-	10.1	25.4	50.5	-
SAM-CLIP [90] <i>ArXiv'23</i>	M. 41M	-	-	17.1	29.2	60.6	-
SimBaseline [57] <i>ECCV'22</i>	COCO S.	-	-	15.3	-	74.5	-
ZegFormer [48] <i>CVPR'22</i>	COCO S.	-	-	16.4	-	73.3	-
LSeg+ [53] <i>CVPR'23</i>	COCO S.	3.8	7.8	18.0	46.5	-	-
X-Decoder [91] <i>CVPR'23</i>	COCO P. + C.	-	-	25.0	-	-	-
OpenSEED [92] <i>ICCV'23</i>	COCO P. + O365	-	-	22.9	-	-	-
MaskCLIP [58] <i>ICML'23</i>	COCO P.	8.2	10.0	23.7	45.9	-	-
OVSeg [55] <i>CVPR'23</i>	COCO S.	9.0	12.4	29.6	55.7	-	94.5
SAN [36] <i>CVPR'23</i>	COCO S.	13.7	17.1	33.3	60.2	-	95.5
OpenSeg [77] <i>ECCV'22</i>	COCO P. + C.	6.3	9.0	21.1	42.1	-	-
ODISE [59] <i>CVPR'23</i>	COCO P.	11.1	14.5	29.9	57.3	84.6	-
FC-CLIP [61] <i>NeurIPS'23</i>	COCO P.	14.8	18.2	34.1	58.4	81.8	95.4
<b>GBA (Ours)</b>	COCO P.	<b>15.1</b>	<b>18.5</b>	<b>35.9</b>	<b>59.6</b>	<b>84.5</b>	<b>95.8</b>



Fig. 6. **Qualitative Visualization of Open-Vocabulary Semantic Segmentation.** To showcase the open-vocabulary recognition capability, we amalgamated class names from all datasets totaling approximately all classes, and conducted open-vocabulary inference directly. The second row represents the segmentation results of the baseline FC-CLIP, and the third row represents the segmentation results of our GBA.

object classification. For instance, in (c), we correctly identify the ‘bear’ instead of misclassifying it as a ‘cat’. This can be attributed to our method’s enhancement of feature diversity and suppression of spurious correlations.

3) *ADE20K Training and COCO Evaluation:* To further substantiate the efficacy of our proposed GBA approach, we conducted experiments utilizing a different training dataset. Specifically, following the methodology of [59], we trained our model on the ADE20K dataset [80] with panoptic annotations and evaluated its performance on the COCO panoptic dataset [79]. As illustrated in Table III, even in this distinct setting (trained on ADE20K and zero-shot evaluated on COCO), our GBA method significantly outperformed the previous state-of-the-art methods, including FreeSeg [60], ODISE [59], and FC-CLIP [61], on the COCO dataset, achieving PQ improvements of 10.5, 2.0, and 2.0 percentage points, respectively. Notably,

TABLE III  
**RESULTS OF TRAINING ON ADE20K PANOPTIC AND EVALUATING ON COCO PANOPTIC VAL SET. THE PROPOSED GBA PERFORMS BETTER THAN PRIOR ARTS, EVEN IN THE DIFFERENT SETTING (TRAINED ON ADE20K AND ZERO-SHOT EVALUATED ON COCO).**

Method	Zero-Shot Test Dataset			Training Dataset		
	COCO			ADE20K		
	PQ	SQ	RQ	PQ	SQ	RQ
FreeSeg [60]	16.5	72.0	21.6	-	-	-
ODISE [59]	25.0	79.4	30.4	31.4	77.9	36.9
FC-CLIP [61]	27.0	78.0	32.9	41.9	78.2	50.2
<b>GBA (Ours)</b>	<b>28.2</b>	<b>78.8</b>	<b>34.1</b>	<b>42.8</b>	<b>80.5</b>	<b>51.3</b>

although our model achieved a slightly lower semantic quality (SQ) score of 1.4 compared to ODISE [59], which employs

a significantly larger backbone network and thus a more robust mask generator, GBA still exhibited superior overall performance due to its simple yet effective design. This further corroborates the robustness and generalization capability of our approach, demonstrating its ability to maintain outstanding open-vocabulary segmentation performance even when applied to different datasets.

TABLE IV

**RESULTS OF VARIANTS.** EXPERIMENTS WERE CONDUCTED ON PASCAL VOC 21, WITH MIoU AS THE EVALUATION METRIC. SDA+CCA REFERS TO CONCATENATING THE TWO ADAPTERS WITH SDA FIRST AND CCA SECOND, WHILE CCA+SDA INDICATES THE REVERSE ORDER.

Method	Stage 1	Stage 2	Stage 3
Plain	82.0	81.9	82.0
SDA+CCA	81.6	81.2	80.4
CCA+SDA	81.4	80.6	79.9

*4) Ablation Studies:* We first investigate the impact of inserting vanilla adapters at different stages. As shown in Table VI, this approach slightly improves performance over the baseline FC-CLIP [61] method at all stages, reaching a maximum mIoU of 82.0. Incorporating SDA at the first stage significantly boosts performance, achieving an mIoU of 82.3. This finding suggests that SDA effectively enhances feature diversity in the early stages of the network, benefiting the subsequent learning process. To independently assess the effectiveness of CCA, we conduct experiments without applying SDA. Inserting CCA at the third stage yields the best performance, reaching an mIoU of 82.5, indicating that CCA more effectively suppresses noisy information in the later stages of the network. We also explore the joint impact of SDA and CCA. When SDA is applied at the first stage and CCA at the third stage, the most significant performance improvement is achieved, reaching an mIoU of 84.5. This result not only confirms the individual effectiveness of SDA and CCA but also highlights the advantages of their collaborative application at specific stages of the network. Furthermore, we visualize the impact of SDA and CCA on the visual features output by CLIP. As shown in Figure 7, the frequency component heatmap of vanilla CLIP is primarily concentrated in the low-frequency regions. SDA enhances the style diversity of visual features through frequency-domain decomposition and amplitude modulation without affecting the spatial characteristics of high and low frequencies. After incorporating CCA, the features begin to focus significantly on high-frequency components. This indicates that CCA emphasizes high-frequency detail features relevant to textual semantics by leveraging cross-modal attention and frequency-domain normalization. These observations are consistent with our ablation study analysis presented in Table VI.

### C. Extending GBA to Transformer Architectures

We apply two adapters from GBA (Style Diversification Adapter and Correlation Constraint Adapter) to the Transformer-based CLIP model, specifically the ViT-B/16 architecture used in ZegFormer [48] and ZegCLIP [94]. We

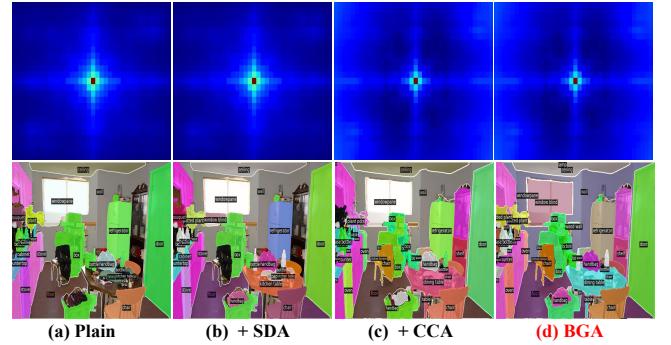


Fig. 7. Visualization of the impact of Style Diversification Adapter (SDA) and Correlation Constraint Adapter (CCA) on frequency components. The changes in the heatmap of high and low-frequency components in the visual features output by CLIP (first row), and their influence on the final segmentation results (second row), after introducing the proposed SDA and CCA to the original CLIP model.

compare the enhanced versions with the original ZegFormer and ZegCLIP. Specifically, we strategically insert the two adapters into the 12 Transformer layers of CLIP: (1) SDA is inserted after the 4th layer to enhance style diversity in the early stages of feature extraction; (2) CCA is placed after the 8th layer to leverage high-level semantic features and suppress spurious correlations. Experimental results (Table V) demonstrate that even with this simple application, GBA achieves improvements on the Transformer architecture. On the PAS-21 dataset, the GBA-enhanced methods outperform the original ZegCLIP and ZegFormer in terms of mIoU for both seen and unseen categories. On the COCO-Stuff dataset, GBA-ZegFormer and GBA-ZegCLIP also exhibit moderate performance gains. Considering that GBA was initially designed for CNN architectures, we believe that with comprehensive experimentation and optimization tailored to the Transformer architecture, GBA has the potential to achieve even greater performance improvements on Transformer-based models.

TABLE V  
**COMPARISON OF TRANSFORMER-BASED METHODS ON PAS-21 AND COCO-STUFF 164K DATASETS.** BEST RESULTS ARE IN BOLD. THESE METHODS DIVIDE THE CLASSES IN EACH DATASET INTO SEEN (S) AND UNSEEN (U) CATEGORIES, TRAINING ONLY ON SEEN CLASSES AND INFERRING ON UNSEEN CLASSES.

Methods	PASCAL VOC 2012			COCO-Stuff 164K		
	pAcc	mIoU(S)	mIoU(U)	pAcc	mIoU(S)	mIoU(U)
ZegFormer	—	86.4	63.6	—	36.6	33.2
<b>+GBA</b>	—	<b>87.3</b>	<b>64.2</b>	—	<b>37.3</b>	<b>33.8</b>
ZegCLIP	94.6	91.9	77.8	62.0	40.2	41.4
<b>+GBA</b>	<b>96.7</b>	<b>92.5</b>	<b>78.8</b>	<b>62.9</b>	<b>40.6</b>	<b>42.0</b>

### D. Exploration of variants of GBA

Table IV shows the experimental results on variants of GBA. The results show that concatenating in this manner negatively impacts performance, as too many features are lost at the same stage, hindering the model's learning process. This effect is more pronounced in deeper stages. Table IV presents the experimental results of different variants of the

proposed method on the PASCAL VOC 21 dataset, using mean Intersection over Union (mIoU) as the evaluation metric. The plain model, without any adapters, serves as the baseline. Two variants, SDA+CCA and CCA+SDA, are investigated, where SDA and CCA adapters are concatenated in different orders. The results demonstrate that concatenating adapters leads to a slight performance degradation compared to the plain model across all three stages. The SDA+CCA variant achieves mIoU scores of 81.6%, 81.2%, and 80.4% in stages 1, 2, and 3, respectively, while the CCA+SDA variant obtains 81.4%, 80.6%, and 79.9% in the corresponding stages. This suggests that the concatenation of adapters may result in some information loss, which hinders the model's learning process. Notably, the performance drop becomes more pronounced in deeper stages, indicating that the impact of information loss accumulates as the network depth increases.

TABLE VI

**ABLATION STUDIES.** Stage 1–3 INDICATES THE LOCATION WHERE THE ADAPTER IS INSERTED. THE ✓ SYMBOL INDICATES THE SELECTED SETTING, WHILE THE ✗ SYMBOL REPRESENTS THE OPPOSITE. w. PLAIN ADAPTER DENOTES THE REPLACEMENT OF SDA AND CCA WITH PLAIN ADAPTERS THAT DOES NOT EMPLOY THE ENHANCED STRATEGIES.

Config.	Stage 1	Stage 2	Stage 3	Method	mIoU
Baseline				FC-CLIP	81.8
w. plain adapter	✓				82.0
		✓			81.9
			✓		82.0
Baseline + SDA	✓				<b>82.3</b>
		✗			82.1
			✗		81.9
Baseline + CCA	✗				82.5
		✗			82.9
			✓		<b>82.5</b>
Baseline + SDA + CCA	✗				83.3
		✗		our GBA	83.8
			✓		<b>84.5</b>

### E. Efficiency Analysis

We analyze the efficiency of our proposed method in terms of computational overhead, parameter count, and inference speed. As shown in Table VII, our method introduces a marginal increase of 3.23% in total parameters and 1.42% in GFLOPs compared to the baseline, indicating minimal computational overhead. Although the number of trainable parameters increases by 58.09%, it primarily originates from the lightweight adapter module and constitutes a small fraction of the overall model size, resulting in a negligible increase in actual inference time. Table VIII presents a frames per second (FPS) comparison, where our method maintains an inference speed comparable to FC-CLIP and significantly outperforms ODISE [59]. On the COCO [79] and Pascal VOC 21 [81] datasets, our approach achieves an FPS of 2.96 and 6.33, respectively. Furthermore, we compare the parameter efficiency and computational cost of our method with other approaches in Table IX. Despite using a higher input resolution and introducing additional parameters through the adapter module, our approach achieves competitive performance while maintaining a reasonable computational cost.

TABLE VII  
COMPARISON OF MODEL COMPLEXITY.

Metric	Baseline	Our Method	Increase (%)
Total Params	372,494,210	384,530,690	3.23%
Trainable Params	20,721,601	32,758,081	58.09%

TABLE VIII  
FPS COMPARISON. FPS COMPARISON OBTAINED USING ONE A6000.

Method	COCO	Pascal VOC 21
ODISE	0.52	0.41
FC-CLIP	3.15	6.45
<b>GBA (Ours)</b>	<b>2.96</b>	<b>6.33</b>

### F. Visualization Assessment on More Challenging Datasets

To further assess the performance of our method on more challenging images, we have selected three representative datasets from the MESS benchmark [96]: the driving imagery dataset BDD-100K [97], the aerial imagery dataset UAvid, and the remote sensing imagery dataset ISPRS Potsdam [98]. Experimental results on BDD-100K (Figure 8 (a),(b) and (c)) demonstrate that GBA exhibits stronger generalization ability on images with different styles compared to the baseline methods. Benefiting from the robustness of the SDA module to image style variations and the enhancement of semantic associations by the CCA module, GBA can more accurately segment target regions under various lighting conditions and effectively suppress background interference. However, the performance on the UAvid [99] and ISPRS Potsdam [98] datasets reveals the challenges faced by cross-domain generalization (Figure 8 (d) and (e)). Aerial and remote sensing images differ significantly from the COCO dataset used for fine-tuning in terms of shooting angle, target scale, and texture features, leading to a larger domain gap between the source dataset and the target domain, which may limit the performance improvement of our method in these specific domains. Despite the challenges in cross-domain generalization, our GBA method demonstrates strong generalization ability on datasets closer to natural images (such as BDD-100K [99]), proving the effectiveness of SDA and CCA in handling the diversity and complexity of natural scenes.

TABLE IX  
COMPARISON OF PARAMETER EFFICIENCY AND COMPUTATIONAL COST ACROSS DIFFERENT METHODS. 'ADAPTER P.' SHOWS THE NUMBER OF PARAMETERS INTRODUCED BY THE ADAPTER MODULE (IF APPLICABLE). 'TRAINABLE P.' REPRESENTS THE TOTAL NUMBER OF TRAINABLE PARAMETERS IN MILLIONS. OUR METHOD USES CONVNeXT AS THE BACKBONE WITH AN INPUT RESOLUTION OF 1024×1024, WHILE OTHER METHODS USE ViT-B/16 WITH 640×640 INPUT RESOLUTION. OVSEG [55] HAS A SIMILAR STRUCTURE TO SIMSEG [57] BUT FINETUNES THE ENTIRE CLIP MODEL, RESULTING IN SIGNIFICANTLY MORE TRAINABLE PARAMETERS. '-' DENOTES UNAVAILABLE DATA.

Method	Backbone	Adapter P.	Trainable P.	GFLOPs
SAN [36]	ViT-B/16	8.4 M	8.4 M	64.3
MaskCLIP [95]	ViT-B/16	—	63.1 M	307.8
SimSeg [57]	ViT-B/16	—	61.1 M	1916.7
OvSeg [55]	ViT-B/16	—	147.2 M	1916.7
FC-CLIP [61]	ConvNeXt-L	—	19.8 M	986.1
<b>BGA (Ours)</b>	ConvNeXt-L	11.5 M	31.3 M	1000.2

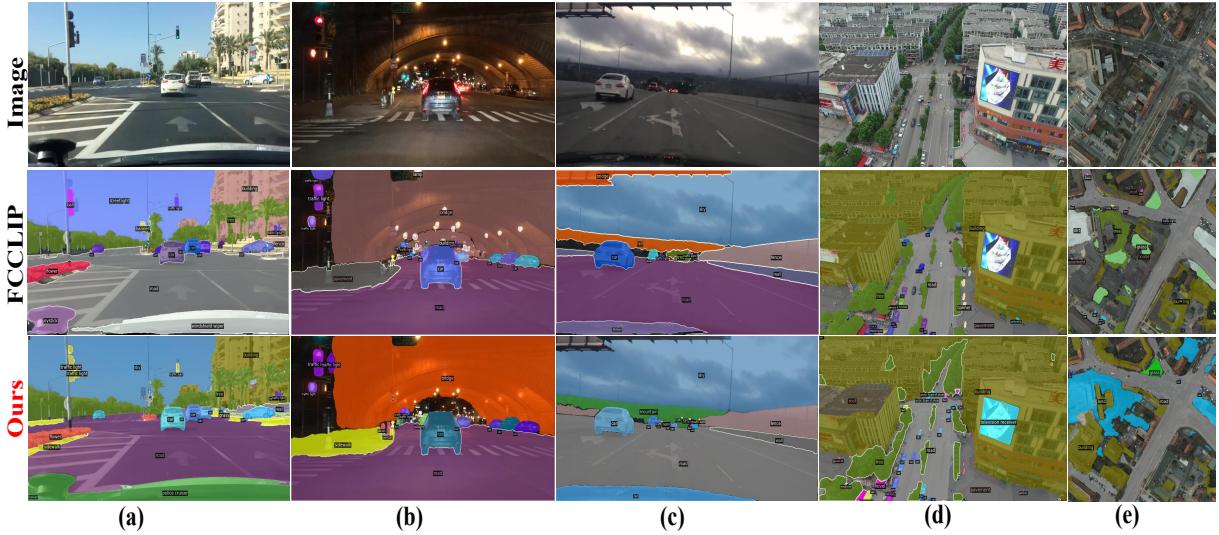


Fig. 8. Open-vocabulary panoptic segmentation performance on the BDD-100K ,UAVid and ISPRS Potsdam datasets. (a), (b) and (c) represent sunny, nighttime, and cloudy scenes, respectively, which represent different image styles and showcase the domain generalization ability of our GBA.

## V. LIMITATIONS

We propose a novel single-stage open-vocabulary segmentation framework that achieves state-of-the-art performance through simple yet effective adapter structures. Despite the encouraging results, we recognize that there are still some limitations and challenges that require further investigation. The main limitations include the need to improve segmentation accuracy for occluded instances and the need to enhance the framework's ability to identify camouflaged instances. These challenges highlight the necessity for further research to improve the robustness and generalization capability of the framework. Furthermore, we identify two interesting future research directions. First, exploring methods to decouple spurious correlations from text features is crucial to prevent the model from being misled by irrelevant information. Second, developing effective techniques to handle conflicting or overlapping vocabulary items, such as distinguishing semantically related but hierarchically different entities (e.g., "dog" and "dog tail"), is an important direction for future research.

## VI. CONCLUSION

The proposed Generalization Boosted Adapter (GBA) strategy effectively enhances the generalization capability and robustness of open-vocabulary segmentation tasks based on the CLIP model. GBA consists of two key components: Style Diversification Adapter (SDA) and Correlation Constraint Adapter (CCA). The SDA, acting on the amplitude component, enriches the feature space representation while preserving content information, thereby mitigating the overfitting problem caused by limited fine-tuning data. The CCA, introduced into the deep layers of the visual encoder, suppresses low-frequency "noise" and improves the accuracy of category matching by avoiding erroneous associations between the correct category and irrelevant features. The synergistic effect between the shallow SDA and deep CCA enables GBA to effectively alleviate overfitting and enhance the semantic

relevance of feature representations. Extensive experiments validate the effectiveness of GBA, demonstrating state-of-the-art performance on multiple benchmarks. The results highlight the potential of GBA as a general solution for improving the generalization capability and robustness of CLIP-based models, paving the way for the design of novel adapter modules in various computer vision tasks.

## REFERENCES

- [1] R. Xu, C. Wang, J. Sun, S. Xu, W. Meng, and X. Zhang, "Self correspondence distillation for end-to-end weakly-supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
- [2] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang, "Wave-like class activation map with representation fusion for weakly-supervised semantic segmentation," *IEEE Transactions on Multimedia*, 2023.
- [3] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Skinformer: Learning statistical texture representation with transformer for skin lesion segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2024.
- [4] B. Wang, X. Gong, P. Lyu, and S. Liang, "Iterative learning-based cooperative motion planning and decision-making for connected and autonomous vehicles coordination at on-ramps," *IEEE Transactions on Intelligent Transportation Systems*, 2024.
- [5] B. Wang, X. Gong, Y. Wang, P. Lyu, and S. Liang, "Coordination for connected and autonomous vehicles at unsignalized intersections: An iterative learning based collision-free motion planning method," *IEEE Internet of Things Journal*, 2023.
- [6] R. Xu, J. Zhang, J. Sun, C. Wang, Y. Wu, S. Xu, W. Meng, and X. Zhang, "Mrftrans: Multimodal representation fusion transformer for monocular 3d semantic scene completion," *Information Fusion*, p. 102493, 2024.
- [7] W. Liu, W. Li, J. Zhu, M. Cui, X. Xie, and L. Zhang, "Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 5855–5867, 2023.
- [8] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.
- [9] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang, "Dc-net: Dual context network for 2d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I* 24. Springer, 2021, pp. 503–513.

- [10] Y. Sun, L. Su, S. Yuan, and H. Meng, "Danet: Dual-branch activation network for small object instance segmentation of ship images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6708–6720, 2023.
- [11] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 4904–4916. [Online]. Available: <http://proceedings.mlr.press/v139/jia21b.html>
- [12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: <http://proceedings.mlr.press/v139/radford21a.html>
- [13] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How can we know what language models know?" *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–438, 2020.
- [14] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Auto-prompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*, 2020.
- [15] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5206–5215.
- [16] W. Xu, R. Xu, C. Wang, S. Xu, L. Guo, M. Zhang, and X. Zhang, "Spectral prompt tuning: Unveiling unseen classes for zero-shot semantic segmentation," *arXiv preprint arXiv:2312.12754*, 2023.
- [17] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.
- [18] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
- [19] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision (IJCV)*, 2022.
- [20] C. Ma, Y. Liu, J. Deng, L. Xie, W. Dong, and C. Xu, "Understanding and mitigating overfitting in prompt tuning for vision-language models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 9, pp. 4616–4629, 2023.
- [21] T. Chang, X. Yang, X. Luo, W. Ji, and M. Wang, "Learning style-invariant robust representation for generalizable visual instance retrieval," *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:264492464>
- [22] K. Song, H. Ma, B. Zou, H. Zhang, and W. Huang, "Fd-align: feature discrimination alignment for fine-tuning pre-trained models in few-shot learning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023, pp. 43579–43592.
- [23] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. Springer, 2020, pp. 104–120.
- [24] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXXI*. Springer, 2020, pp. 121–137.
- [25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
- [26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [27] Y. Li, L. Song, Y. Chen, Z. Li, X. Zhang, X. Wang, and J. Sun, "Learning dynamic routing for semantic segmentation," in *CVPR*, 2020.
- [28] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and W. He, "Navid: Video-based vlm plans the next step for vision-and-language navigation," *arXiv preprint arXiv:2402.15852*, 2024.
- [29] D. Zhang, H. Li, W. Cong, R. Xu, J. Dong, and X. Chen, "Task relation distillation and prototypical pseudo label for incremental named entity recognition," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 2023, pp. 3319–3329.
- [30] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.
- [31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [32] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang, "Realm: Retrieval-augmented language model pre-training," *ArXiv*, vol. abs/2002.08909, 2020. [Online]. Available: <https://api.semanticscholar.org/CorpusID:211204736>
- [33] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16795–16804, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:247363011>
- [34] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. J. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *ArXiv*, vol. abs/2110.04544, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:238583492>
- [35] T. Chen, L. Zhu, C. Deng, R. Cao, Y. Wang, S. Zhang, Z. Li, L. Sun, Y. Zang, and P. Mao, "Sam-adapter: Adapting segment anything in underperformed scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3367–3375.
- [36] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "Side adapter network for open-vocabulary semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2945–2954.
- [37] R. Xu, Y. Li, C. Wang, S. Xu, W. Meng, and X. Zhang, "Instance segmentation of biological images using graph convolutional network," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104739, 2022.
- [38] C. Wang, R. Xu, S. Xu, W. Meng, and X. Zhang, "Cndesc: Cross normalization for local descriptors learning," *IEEE Transactions on Multimedia*, 2022.
- [39] S. Xu, S. Chen, R. Xu, C. Wang, P. Lu, and L. Guo, "Local feature matching using deep learning: A survey," *arXiv preprint arXiv:2401.17592*, 2024.
- [40] R. Xu, C. Wang, S. Xu, W. Meng, Y. Zhang, B. Fan, and X. Zhang, "Domainfeat: Learning local features with domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [41] J. Ji, R. Shi, S. Li, P. Chen, and Q. Miao, "Encoder-decoder with cascaded crfs for semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1926–1938, 2020.
- [42] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation," *IEEE Transactions on Image Processing*, vol. 32, pp. 1052–1064, 2023.
- [43] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang, "Dual-stream representation fusion learning for accurate medical image segmentation," *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106402, 2023.
- [44] S. Xu, S. Zheng, W. Xu, R. Xu, C. Wang, J. Zhang, X. Teng, A. Li, and L. Guo, "Hcf-net: Hierarchical context fusion network for infrared small object detection," *arXiv preprint arXiv:2403.10778*, 2024.
- [45] C. Wang, R. Xu, S. Xu, W. Meng, and X. Zhang, "Treating pseudo-labels generation as image matting for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 755–765.
- [46] J. Chen, W. Lu, Y. Li, L. Shen, and J. Duan, "Adversarial learning of object-aware activation map for weakly-supervised semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
- [47] W. Liu, G. Lin, T. Zhang, and Z. Liu, "Guided co-segmentation network for fast video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1607–1617, 2020.
- [48] J. Ding, N. Xue, G.-S. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *CVPR*, 2022.
- [49] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *European Conference on Computer Vision*, 2022, pp. 736–753.

- [50] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic projection network for zero-and few-label semantic segmentation," in *CVPR*, 2019.
- [51] M. Bucher, T.-H. Vu, M. Cord, and P. Pérez, "Zero-shot semantic segmentation," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 468–479.
- [52] G. Pastore, F. Cermelli, Y. Xian, M. Mancini, Z. Akata, and B. Caputo, "A closer look at self-training for zero-label semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2693–2702.
- [53] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *ICLR*, 2022.
- [54] D. Huynh, J. Kuen, Z. Lin, J. Gu, and E. Elhamifar, "Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7020–7031.
- [55] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7061–7070.
- [56] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *European Conference on Computer Vision*. Springer, 2022, pp. 696–712.
- [57] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *European Conference on Computer Vision*. Springer, 2022, pp. 736–753.
- [58] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary panoptic segmentation with maskclip," *arXiv preprint arXiv:2208.08984*, 2022.
- [59] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. 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*, 2023, pp. 2955–2966.
- [60] J. Qin, J. Wu, P. Yan, M. Li, R. Yuxi, X. Xiao, Y. Wang, R. Wang, S. Wen, X. Pan *et al.*, "Freeseg: Unified, universal and open-vocabulary image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19446–19455.
- [61] Q. Yu, J. He, X. Deng, X. Shen, and L.-C. Chen, "Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip," *arXiv preprint arXiv:2308.02487*, 2023.
- [62] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017, pp. 2961–2969.
- [63] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, "Panoptic segmentation," in *CVPR*, 2019, pp. 9404–9413.
- [64] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *CVPR*, 2022.
- [65] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *ArXiv*, vol. abs/1502.03167, 2015.
- [66] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [67] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *ArXiv*, vol. abs/1607.08022, 2016.
- [68] X. Jin, C. Lan, W. Zeng, Z. Chen, and L. Zhang, "Style normalization and restitution for generalizable person re-identification," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3140–3149, 2020.
- [69] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *ECCV*, 2018.
- [70] G. Chen, P. Peng, L. Ma, J. Li, L. Du, and Y. Tian, "Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 448–457, 2021.
- [71] S. Lee, J. Bae, and H. Y. Kim, "Decompose, adjust, compose: Effective normalization by playing with frequency for domain generalization," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11776–11785, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:257364773>
- [72] Z. Wang, Y. Luo, Z. Huang, and M. Baktashmotagh, "Ffm: Injecting out-of-domain knowledge via factorized frequency modification," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4124–4133, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:256660128>
- [73] J. Bai, L. Yuan, S.-T. Xia, S. Yan, Z. Li, and W. Liu, "Improving vision transformers by revisiting high-frequency components," in *European Conference on Computer Vision*. Springer, 2022, pp. 1–18.
- [74] Y.-Q. Sun, J.-W. Tian, and J. Liu, "Background suppression based-on wavelet transformation to detect infrared target," in *2005 International Conference on Machine Learning and Cybernetics*, vol. 8. IEEE, 2005, pp. 4611–4615.
- [75] J. Liu, J. Gao, Q. Shen, and J. Tian, "A background separation method of nonuniform image segmentation," in *2009 4th IEEE Conference on Industrial Electronics and Applications*. IEEE, 2009, pp. 3049–3053.
- [76] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," in *NeurIPS*, 2022.
- [77] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European Conference on Computer Vision*, 2022, pp. 540–557.
- [78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
- [79] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
- [80] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *CVPR*, 2017, pp. 633–641.
- [81] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, pp. 303–338, 2010.
- [82] X. Wang, R. Girdhar, S. X. Yu, and I. Misra, "Cut and learn for unsupervised object detection and instance segmentation," in *CVPR*, 2023, pp. 3124–3134.
- [83] D. Niu, X. Wang, X. Han, L. Lian, R. Herzig, and T. Darrell, "Unsupervised universal image segmentation," *arXiv preprint arXiv:2312.17243*, 2023.
- [84] J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang, "Groupvit: Semantic segmentation emerges from text supervision," in *CVPR*, 2022.
- [85] J. Cha, J. Mun, and B. Roh, "Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs," in *CVPR*, 2023, pp. 11165–11174.
- [86] J. Xu, J. Hou, Y. Zhang, R. Feng, Y. Wang, Y. Qiao, and W. Xie, "Learning open-vocabulary semantic segmentation models from natural language supervision," in *CVPR*, 2023, pp. 2935–2944.
- [87] H. Luo, J. Bao, Y. Wu, X. He, and T. Li, "Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," in *ICML*, 2023, pp. 23033–23044.
- [88] K. Ranasinghe, B. McKinzie, S. Ravi, Y. Yang, A. Toshev, and J. Shlens, "Perceptual grouping in contrastive vision-language models," in *ICCV*, 2023, pp. 5571–5584.
- [89] K. Cai, P. Ren, Y. Zhu, H. Xu, J. Liu, C. Li, G. Wang, and X. Liang, "Mixreorg: Cross-modal mixed patch reorganization is a good mask learner for open-world semantic segmentation," in *ICCV*, 2023, pp. 1196–1205.
- [90] H. Wang, P. K. A. Vasu, F. Faghri, R. Vemulapalli, M. Farajtabar, S. Mehta, M. Rastegari, O. Tuzel, and H. Pouransari, "Sam-clip: Merging vision foundation models towards semantic and spatial understanding," *arXiv preprint arXiv:2310.15308*, 2023.
- [91] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, H. Behl, J. Wang, L. Yuan *et al.*, "Generalized decoding for pixel, image, and language," in *CVPR*, 2023, pp. 15116–15127.
- [92] H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Yang, and L. Zhang, "A simple framework for open-vocabulary segmentation and detection," in *ICCV*, 2023, pp. 1020–1031.
- [93] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, "The role of context for object detection and semantic segmentation in the wild," in *CVPR*, 2014.
- [94] Z. Zhou, Y. Lei, B. Zhang, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11175–11185.
- [95] Z. Ding, J. Wang, and Z. Tu, "Open-vocabulary universal image segmentation with maskclip," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 8090–8102.
- [96] B. Blumenstiel, J. Jakubik, H. Kuhne, and M. Vossing, "What a mess: Multi-domain evaluation of zero-shot semantic segmentation," *ArXiv*, vol. abs/2306.15521, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:259261907>
- [97] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

- [98] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:248157640>
- [99] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 165, pp. 108–119, 2020.