

DPSeg: Dual-Prompt Cost Volume Learning for Open-Vocabulary Semantic Segmentation

Ziyu Zhao^{1*}, Xiaoguang Li^{1*}, Linjia Shi¹, Nasrin Imanpour¹, Song Wang^{2†}

¹University of South Carolina, USA ²Shenzhen University of Advanced Technology, China

{ziyuz, xl22, linjia, imanpour}@email.sc.edu, wangsong@suat-sz.edu.cn

Abstract

Open-vocabulary semantic segmentation aims to segment images into distinct semantic regions for both seen and unseen categories at the pixel level. Current methods utilize text embeddings from pre-trained vision-language models like CLIP but struggle with the inherent domain gap between image and text embeddings, even after extensive alignment during training. Additionally, relying solely on deep text-aligned features limits shallow-level feature guidance, which is crucial for detecting small objects and fine details, ultimately reducing segmentation accuracy. To address these limitations, we propose a dual prompting framework, DPSeg, for this task. Our approach combines dual-prompt cost volume generation, a cost volume-guided decoder, and a semantic-guided prompt refinement strategy that leverages our dual prompting scheme to mitigate alignment issues in visual prompt generation. By incorporating visual embeddings from a visual prompt encoder, our approach reduces the domain gap between text and image embeddings while providing multi-level guidance through shallow features. Extensive experiments demonstrate that our method significantly outperforms existing state-of-the-art approaches on multiple public datasets.

1. Introduction

Semantic segmentation, the process of assigning class labels to individual pixels in an image, has seen significant advancements. Conventional segmentation methods [4, 5, 21, 25, 28, 42, 64, 65] are often structured for closed-set tasks, wherein models are trained and evaluated on a fixed set of categories. While these models demonstrate strong performance in controlled environments, they typically struggle to generalize effectively to real-world scenarios that include unseen objects and novel categories.

Open-vocabulary semantic segmentation (OVSS) has

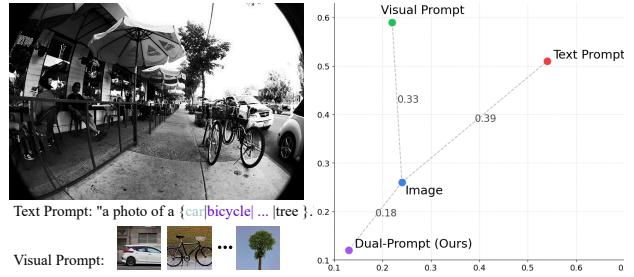


Figure 1. **T-SNE visualization of embeddings in CLIP feature space, showing text prompt, visual prompt, target image, and our dual-prompt for a street scene image.** Numbers indicate distance to image feature, where our dual-prompt approach achieves closest proximity (0.18).

emerged as a promising solution to this limitation. By leveraging large-scale vision-language models like CLIP [37], OVSS enables segmentation beyond fixed categories, allowing models to generalize to arbitrary classes during inference. Recent approaches, such as CAT-Seg, [7] utilize pixel-level cost volumes—dense similarity mappings between local image features and text embeddings in CLIP’s space—to balance segmentation accuracy and efficiency. SED [52] further enhances this framework by incorporating a hierarchical encoder and multi-scale features. However, existing methods face two key challenges: (1) despite incorporating multi-scale semantic information, the reliance on upsampling the cost volume leads to potential degradation of fine-grained details, particularly affecting small but crucial semantic segments, and (2) the sole dependence on text-image embedding alignment fails to fully resolve the domain gap, even with robust pre-trained models.

To overcome these limitations, our work begins by conducting comprehensive experiments (see Sec. 3.1 and Sec. 3.2) to investigate the domain gap between image and text embeddings by calculating cosine similarity scores across sampled categories. Our findings reveal that visual prompts exhibit stronger feature alignment with image representations in the embedding space compared to text prompts. Additionally, the cost volumes generated from vi-

*Co-first authors and contribute equally.

†Corresponding author.

sual prompts demonstrate richer spatial-semantic cues for precise segment localization than their text-based counterparts. Building on these insights, we propose a novel OVSS framework with two key components: (1) a dual-prompt cost volume generation mechanism that leverages both visual and text prompts, and (2) a cost volume-guided decoder that progressively integrates visual prompt features at corresponding feature levels for refined segmentation prediction. To further improve segmentation accuracy and mitigate potential misalignment between visual prompts and actual segments during inference, we propose a semantic-guided prompt refinement strategy that employs a two-stage process, where initial segmentation results serve as refined visual prompts to guide the final prediction. As visualized in Fig. 1, our dual-prompt scheme achieves superior embedding space alignment with a distance of 0.18 to image features, compared to 0.33 for visual prompts and 0.39 for text prompts alone. The t-SNE visualization demonstrates that our dual-prompt approach effectively bridges the semantic gap by integrating synergistic information from both modalities. Extensive experiments show that our method sets a new benchmark in open-vocabulary semantic segmentation, achieving superior accuracy, inference efficiency, and overall performance. Our main contribution is the following:

- Through both quantitative and qualitative analysis of CLIP embeddings, we demonstrate that visual prompts achieve stronger feature alignment and generate more precise spatial-semantic cost volumes compared to text-based approaches.
- We propose **DPSeg**, a novel approach for open-vocabulary semantic segmentation featuring dual-prompt cost volume generation and cost volume-guided decoder. Additionally, our semantic-guided prompt refinement module leverages initial predictions as scene-specific visual prompts to enhance segmentation quality through a two-stage inference process.
- The extensive experiment results on multiple open vocabulary semantic segmentation datasets demonstrate that our method outperforms competitors by a large margin.

2. Related Work

2.1. Cross-Modality Learning Models

Cross-modality representation learning has seen significant advancements, broadly categorized into Spatial Cross Modality (2D-3D) [65] and Conceptual Cross Modality (Vision-Language). Vision-language models [6, 29, 39, 43] have been developed and fine-tuned on various downstream tasks using image-text pairs. CLIP [37, 58], for instance, leverages large-scale image-text data to align visual and language representations, excelling in zero-shot image recognition and expanding into tasks like segmentation [15, 23, 54, 56], captioning [30], classification [1], restaura-

tion [60], and object detection [13, 66]. Recent CLIP-based extensions have shown notable improvements across various tasks. SLIP [32] enhances CLIP by incorporating self-supervised image-to-image contrastive learning alongside CLIP’s image-text alignment, yielding richer and more robust visual representations that capture subtle image features more effectively. A-CLIP [57] introduces an attentive token removal strategy, retaining only a small subset of tokens that exhibit strong semantic correlations with the corresponding text descriptions, allowing for more focused and relevant visual-text alignment.

In segmentation, OVSeg [23] adapts CLIP by fine-tuning on masked image regions to produce mask-aware embeddings for object-level segmentation. However, it struggles to capture detailed semantic information, such as object attributes and relationships, reducing its effectiveness in complex scenes. Our model addresses this by leveraging higher intra-modal similarity through integrated visual prompts, enriching cross-modal cosine similarity and supplementing semantic information beyond text prompts alone.

2.2. Open-Vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation (OVSS) enables models to segment arbitrary object categories, including unseen classes. While early zero-shot approaches relied on attribute-based classifiers [22] and word embeddings [63], vision-language models like CLIP [37] and ALIGN [18] have significantly advanced the field through image-text alignment in shared embedding spaces. OVSS methods generally follow either ① a two-stage pipeline [8, 9, 12, 17, 23, 55] or ② an end-to-end network [53, 54, 56, 69]. Among two-stage approaches, SCAN [27] calibrates CLIP with semantic priors, EBSeg [40] employs balanced decoders with semantic consistency, and USE [48] leverages multi-granularity segments. End-to-end methods have progressed from MaskCLIP’s [68] direct CLIP adaptation to SAN’s [56] sophisticated side-adapter network. Recent advances like CAT-Seg [7] and SED [52] utilize cost volumes with auxiliary backbones and hierarchical feature fusion, respectively. Building upon these works, we address the limitation of detail degradation during cost volume upsampling by proposing a multi-scale approach that combines both visual and text prompts to better utilize feature information across scales.

2.3. Visual Prompting

Prompting generally refers to providing textual input to an AI model to elicit a specific task. Visual prompting extends this concept by using an image as a prompt, where semantically related reference images guide the model in tasks such as segmentation [65]. Visual prompting can be broadly categorized into prototype-based and feature-matching approaches. Prototype-based methods, such as PFENet [68]

and PANet [46], extract average prototypes of embeddings for each semantic category. Feature-matching methods, like CyCTR [61] and HDMNet [35], use pixel-level correlations between reference and target images to boost segmentation performance. In recent segmentation advancements, GFSS [16] uses learned visual prompts with a transformer decoder to extract prototypes for generalized few-shot segmentation. We employ Stable Diffusion [38] for automatic visual prompt generation aligned with text descriptions, enabling CLIP to produce rich semantic cost volumes through dual-prompt integration.

3. Discussion and Motivation

In this section, we conduct experiments to analyze the modality gap [19, 24] between the aligned image embeddings and text embeddings extracted by pre-trained vision-language models, such as CLIP, as described in Sec. 3.1. Additionally, we explore cost volume visualization to highlight the benefits of visual prompts by comparing the cost volumes generated from text and visual prompts, as shown in Sec. 3.2. Based on the above analysis and observations, we propose combining visual and text prompts to enhance segmentation accuracy, particularly in scenarios involving both seen and unseen categories.

3.1. Modality Gap of the Aligned Image and Text Embeddings

To demonstrate the modality gap [19, 24] between aligned image and text embeddings, we conduct the experiment by following these steps: ① Given an image \mathbf{I} and its corresponding text prompt \mathcal{P}_t , we use the pre-trained CLIP model to extract their embeddings \mathbf{E} and \mathbf{T} , respectively. ② Using the pre-trained Stable Diffusion model [38] to generate the synthetic visual prompt \mathcal{P}_v by inputting \mathcal{P}_t and then extract its corresponding embedding \mathbf{V} using the same CLIP image encoder. ③ Calculating the cosine similarity between \mathbf{E} and \mathbf{T} , as well as between \mathbf{E} and \mathbf{V} , across more than one hundred samples. As shown in Fig. 2, the similarity between \mathbf{E} and \mathbf{V} is significantly higher than that between \mathbf{E} and \mathbf{T} , demonstrating that the visual prompt \mathcal{P}_v , which shares the same modality as \mathbf{I} , can significantly alleviate the embedding modality gap.

3.2. Effectiveness of the Visual Prompt in Semantic Segmentation

To analyze the effectiveness of the visual prompt in semantic segmentation, we conduct the experiment as follows: ① Using the pre-trained CLIP image encoder, we extract the image feature from the layer before pooling. ② Using the same approach as in Sec. 3.1, we extract the visual prompt embedding \mathbf{V} and the text prompt embedding \mathbf{T} . ③ We then calculate the similarity between the image feature of each pixel location and the visual prompt embedding \mathbf{V} as well

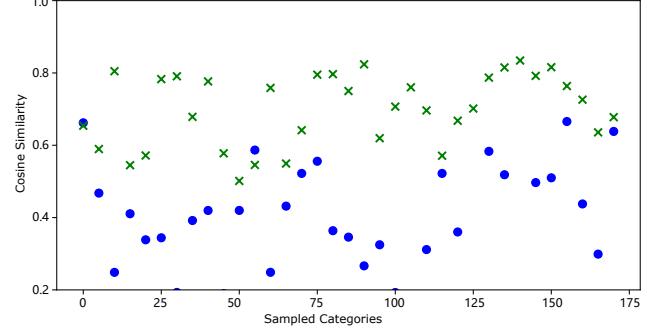


Figure 2. Visualization of cosine similarities comparing image embeddings \mathbf{E} with text prompt embeddings \mathbf{T} (blue dots) and visual prompt embeddings \mathbf{V} (green crosses) across sampled categories.

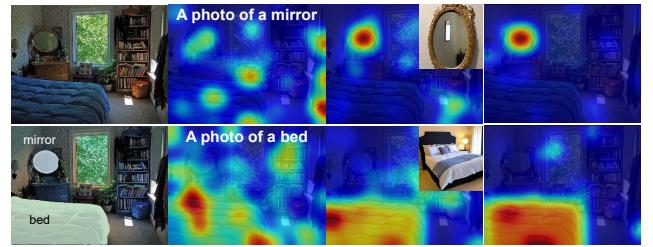


Figure 3. Visualization of cost volume: (a) image with ‘mirror’ and ‘bed’ segments; (b) cost volume with text prompts; (c) cost volume with visual prompts; (d) aggregated cost volume. The top row represents the unseen class ‘mirror,’ and the bottom row represents the seen class ‘bed’.

as the text prompt embedding \mathbf{T} . As shown in the heatmap in Fig. 3, visual prompts yield sharper and more accurate semantic contours than text prompts, particularly for details like “mirror” and “bed” in Fig. 3 (b) and (c).

Inspired by this observation, in this work we propose to leverage both visual and text prompts to enhance the segmentation performance. Specifically, we average the visual and text prompt embeddings and calculate the cosine similarity as described above. As illustrated in Fig. 3 (d), using both visual and text prompts reduces noise and captures more accurate semantic information.

4. Method

Building on insights from Sec. 3, we propose our novel framework **DPSeg** for open-vocabulary semantic segmentation, comprising three core components: a dual-prompt cost volume generation module (see Sec. 4.1), a cost volume-guided decoder (see Sec. 4.2), and a semantic-guided prompt refinement module (see Sec. 4.3), as shown in Fig. 4 and Fig. 5. The dual-prompt cost volume generation module utilizes both text and visual prompts to generate a pixel-level cost volume \mathcal{F}_c with image features. The cost volume-guided decoder integrates multi-scale cost volumes $\mathcal{F}_v^2, \mathcal{F}_v^3, \mathcal{F}_v^4$ from the image and visual prompts to

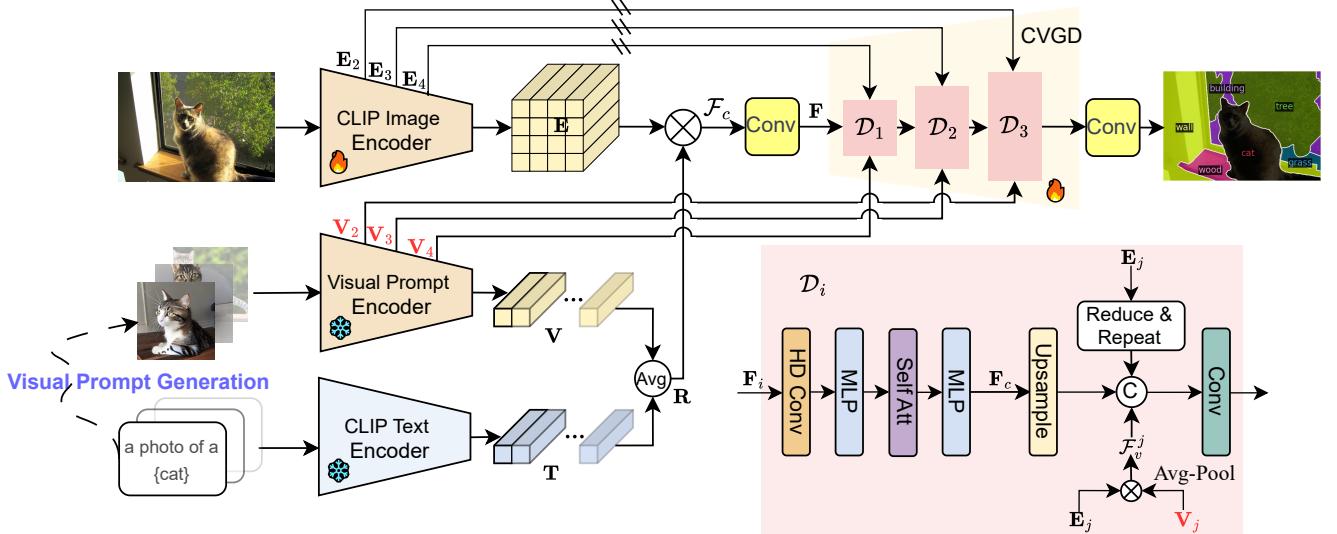


Figure 4. Architecture of our DPSeg network. We begin by generating visual prompts based on text prompt templates, which are then combined with text prompts to create the dual-prompt cost volume. Subsequently, we incorporate multi-scale image features and corresponding visual prompt embeddings into the cost volume-guided decoder (CVGD) in a progressive manner.

guide segmentation map prediction. To address misalignment and uncertainty in visual prompt generation, we propose a two-pass inference strategy, semantic-guided prompt refinement, to enhance alignment and improve segmentation accuracy.

4.1. Dual-Prompt Cost Volume Generation

Prior open-vocabulary semantic segmentation approaches [7, 52] mainly rely on text-image feature alignment using CLIP to generate pixel-level cost volumes. However, such cross-modal alignment can be limited in precision due to inherent differences between text and image representations, even with advanced pre-training as discussed in Sec. 3. To address this, we introduce a cross-modal text-visual prompt cost volume generation, which enhances the alignment within the visual domain by introducing higher intra-modality similarity.

Image encoder. For an input RGB image \mathbf{I} , we employ a hierarchical encoder to extract image features $\mathbf{E}_j, j \in (2, 3, 4, 5)$ at progressively coarser resolutions, corresponding to scales of $4\times, 8\times, 16\times$, and $32\times$ smaller than the original input size. The final layer, \mathbf{E}_5 , is processed through an MLP layer to align the dimensionality with the prompt embeddings, yielding the final image embedding $\mathbf{E} \in \mathbb{R}^{H \times W \times D_z}$, where H , W , and D_z represent the height, width, and number of channels respectively.

Text prompt embedding generation. For generating text prompt embeddings, we adopt a diverse set of descriptive templates following the strategies in [7, 13, 23, 52]. These templates provide varied descriptions for each category C_k , such as “a photo of a $\{C_k\}$ ”, “a $\{C_k\}$ in the scene”, etc. Each description $P_t^{(k,m)}$ is processed through

CLIP’s text encoder, producing text embeddings $\mathbf{T} = \{\mathbf{T}_{k,m} \mid k = 1, \dots, K; m = 1, \dots, M\} \in \mathbb{R}^{K \times M \times D_z}$, where K is the total number of categories, M is the number of templates per category, and D_z is the dimension of the text prompt embeddings, which matches that of the image embeddings for consistency. This variety of templates enhances the semantic richness of each category representation, supporting robust category detection across diverse contexts.

Visual prompt embedding generation. For visual prompt generation, we employ a pre-trained Stable Diffusion model [38] to generate visual prompts by inputting text prompt $\mathcal{P}_t^{(k,m)}$, with each template producing a distinct prompt prototype for the category. The generated visual prompts are processed through a visual prompt encoder, structurally identical to the image encoder. The visual prompt encoder produces an embedding $\mathbf{V} \in \mathbb{R}^{K \times M \times D_z}$, aligned with the text prompt embeddings for seamless integration. Additionally, we extract multi-scale features $\mathbf{V}_j \in \mathbb{R}^{H_j \times W_j \times K \times M \times D_j}$ (for $j = 2, 3, 4$), where H_j and W_j denote the spatial dimensions downsampled by 4, 8 and 16, respectively; K and M represent the category and prototypes of visual prompts, and D_j is the feature dimension at each level j . These multi-scale features further enrich the cost volume, providing fine-grained semantic information that enhances decoding in the segmentation process.

Combined Cost Volume Generation. To utilize both visual and text prompt embeddings, we average them to create a unified representation \mathbf{R} , denoted as

$$\mathbf{R} = \text{Avg}(\mathbf{V} + \mathbf{T}).$$

Then we calculate the dual-prompt cost volume $\mathcal{F}_c \in$

$\mathbb{R}^{H \times W \times K \times M}$ using the \mathbf{R} and image feature \mathbf{E} , as follows:

$$\mathcal{F}_c(x, y, k, m) = \frac{\mathbf{E}(x, y) \cdot \mathbf{R}(k, m)}{\|\mathbf{E}(x, y)\| \|\mathbf{R}(k, m)\|}, \quad (1)$$

where (x, y) represents the spatial location, and (k, m) denotes the category and template index, respectively. To facilitate efficient processing of this high-dimensional cost volume, we apply a convolution layer independently to each cost slice, producing an initial cost volume embedding. This embedding is then fed into the decoder as $\mathbf{F} \in \mathbb{R}^{(H \times W) \times K \times d_F}$, where d_F is the cost volume embedding dimension.

4.2. Cost Volume-Guided Decoder

Building on the findings in Sec. 3, we design a Cost Volume-Guided Decoder that leverages dual-prompt cost volume and hierarchical encoder structure. At each decoder stage \mathcal{D}_i ($i \in (1, 2, 3)$), the input feature map \mathbf{F}_i is processed through a hybrid dilated convolution (HD Conv) module with a 3×3 kernel and dilation rates of 1, 2, and 4, providing large receptive fields (3×3 , 5×5 , and 9×9) with fewer parameters than standard large-kernel convolutions. These features are merged via element-wise addition, followed by normalization, an MLP with GeLU activation, and a self-attention layer for category-level aggregation.

To integrate multi-scale image features from the encoder, which captures rich local details, we first upsample the feature map \mathbf{F}_c following the self-attention layer by a factor of 2 using deconvolution. Each scale feature \mathbf{E}_j , $j \in (2, 3, 4)$ is then reduced by a factor of 16 through convolution, repeated the number of categories N times to align with the category dimension of the upsampled feature map, and finally concatenated with it. As suggested in [7], we avoid back-propagating these intermediate features directly to the image encoder to prevent potential performance degradation.

Instead of upsampling initial cost volume, we employ intra-modal similarity by integrating intermediate-scale image features $\mathbf{E}_j \in \mathbb{R}^{K \times M \times D_j}$, $j \in (2, 3, 4)$, with corresponding multi-scale visual prompt features $\mathbf{V}_j \in \mathbb{R}^{H_j \times W_j \times K \times M \times D_j}$, $j \in (2, 3, 4)$ to compute the visual cost volumes, thereby enhancing semantic coherence across scales. This method captures fine-grained semantic information at multiple levels, yielding visual cost volumes \mathcal{F}_v^j at each stage, calculated as:

$$\mathcal{F}_v^j(x, y, k, m) = \frac{Pool(\mathbf{V}_j(k, m)) \cdot \mathbf{E}_j(x, y)}{\|Pool(\mathbf{V}_j(k, m))\| \|\mathbf{E}_j(x, y)\|}. \quad (2)$$

Here, global average pooling is applied along the H_j and W_j . Each channel in the visual cost volume directly aligns with the corresponding decoding feature channels at each decoder stage, eliminating the need for additional upsampling and preserving both semantic and spatial details. The

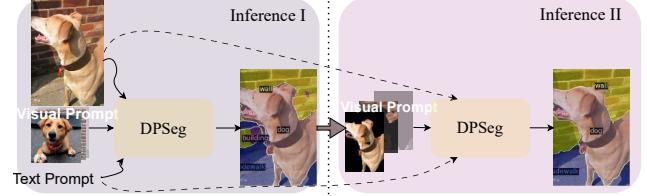


Figure 5. Structure of Semantic-Guided Prompt Refinement.

final high-dimensional feature map is then mapped to the segmentation output through a convolution layer, enabling accurate predictions across all semantic categories for open-vocabulary semantic segmentation.

4.3. Semantic-Guided Prompt Refinement

The visual prompt generation process during training and initial inference introduces variability, as visual prompts generated from text prompts may not fully align with the input image \mathbf{I} . This misalignment can impact segmentation accuracy, particularly for fine-grained details or objects that are not well-represented by the initial visual prompt. To address this challenge, we propose a semantic-guided prompt refinement inference strategy, illustrated in Fig. 5. The first inference (Inference I) pass employs text prompts and initial visual prompts to produce preliminary segmentation results. For each category detected in the first pass, its corresponding segmentation mask is used to crop the objects, which will replace all visual prompts of that category in the second inference pass (Inference II) while maintaining initial prompts for undetected categories and text prompts. This instance-aware refinement enhances the cost volume’s semantic alignment with actual objects in the current image, achieving more precise object boundaries and local details.

5. Experiments

5.1. Setups

Datasets. To acquire object information and evaluate our models, we follow a dataset setup similar to that in [52]. We train our model on the **COCO-Stuff** [3] dataset and evaluate its performance across five additional benchmarks: **Pascal VOC** [11], **Pascal Context** [31] (referred to as *PC-59* and *PC-459*), **ADE20K-150** (referred to as *A-150*) [67], and **ADE20K-847** (referred to as *A-847*) [67]. **COCO-Stuff** is a large-scale semantic segmentation dataset comprising 171 densely annotated classes, with 118k training images, 5k validation images, and 41k test images. **Pascal VOC**, a traditional dataset for segmentation and object detection, includes 20 classes, with 1464 training and 1449 validation images. **Pascal Context** is an extension of **Pascal VOC**, with 1.5k training images and 1.5k validation images, offering two types of annotations: *PC-59* (59 classes) and *PC-459* (459 classes). The **ADE20K** dataset consists of 20k

Method	VLM	Backbone	Training Dataset	A-847	PC-459	A-150	PC-59	PAS-20
ZS3Net [2]	-	ResNet-101	PASCAL VOC	-	-	-	19.4	38.3
LSeg [51]	ViT-B/32	ResNet-101	PASCAL VOC-15	-	-	-	-	47.4
LSeg+ [12]	ALIGN RN-101	ResNet-101	COCO-Stuff	2.5	5.2	13.0	36.0	-
Han et al. [14]	ViT-B/16	ResNet-101	COCO Panoptic [20]	3.5	7.1	18.8	45.2	83.2
ZegFormer [8]	ViT-B/16	ResNet-101	COCO-Stuff	5.6	10.4	18.0	45.5	89.5
SimSeg [55]	ViT-B/16	ResNet-101	COCO-Stuff	7.0	-	20.5	47.7	88.4
OpenSeg [12]	ALIGN	ResNet-101	COCO Panoptic [20] + Loc. Narr. [36]	4.4	7.9	17.5	40.1	-
PACL [33]	ViT-B/16	-	GCC [41] + YFCC[45]	-	-	31.4	50.1	72.3
OVSeg [23]	ViT-B/16	ResNet-101c	COCO-Stuff+COCO Caption	7.1	11.0	24.8	53.3	92.6
CAT-Seg [7]	ViT-B/16	ResNet-101	COCO-Stuff	8.4	16.6	27.2	57.5	93.7
SAN [56]	ViT-B/16	-	COCO-Stuff	10.1	12.6	27.5	53.8	94.0
EBSeg [40]	ViT-B/16	-	COCO-Stuff	11.1	17.3	30.0	56.7	94.6
SCAN [27]	ViT-B/16	Swin-B	COCO-Stuff	10.8	13.2	30.8	58.4	97.0
SED [52]	ConvNeXt-B	-	COCO-Stuff	11.4	18.6	31.6	57.3	94.4
DPSeg (Inference I)	ConvNeXt-B	-	COCO-Stuff	<u>12.0</u>	<u>19.5</u>	<u>32.9</u>	<u>58.1</u>	96.0
DPSeg (Inference II, Ours)	ConvNeXt-B	-	COCO-Stuff	12.5	20.1	33.3	58.4	96.9
LSeg [51]	ViT-B/32	ViT-L/16	PASCAL VOC-15	-	-	-	-	52.3
OpenSeg [12]	ALIGN	Eff-B7 [44]	COCO Panoptic [20] + Loc. Narr. [36]	8.1	11.5	26.4	44.8	-
OVSeg [23]	ViT-L/14	Swin-B	COCO-Stuff+COCO Caption	9.0	12.4	29.6	55.7	94.5
Ding et al.	ViT-L/14	-	COCO Panoptic [20]	8.2	10.0	23.7	45.9	-
ODISE [54]	ViT-L/14	-	COCO Panoptic [20]	11.1	14.5	29.9	57.3	-
HIPIE [49]	BERT-B	ViT-H	COCO Panoptic [20]	-	-	29.0	59.3	-
SAN [56]	ViT-L/14	-	COCO-Stuff	13.7	17.1	33.3	60.2	95.5
CAT-Seg [7]	ViT-L/14	Swin-B	COCO-Stuff	10.8	20.4	31.5	62.0	96.6
FC-CLIP [59]	ConvNeXt-L	-	COCO Panoptic [20]	14.8	18.2	34.1	58.4	95.4
EBSeg [40]	ViT-L/14	-	COCO-Stuff	13.7	21.0	32.8	60.2	96.4
SCAN [26]	ViT-L/14	Swin-B	COCO-Stuff	14.0	16.7	33.5	59.3	97.2
USE+SAM [47]	ViT-L/14	Swin-B	COCO-Stuff	13.3	14.7	37.0	57.8	-
SED [52]	ConvNeXt-L	-	COCO-Stuff	13.9	22.6	35.2	60.6	96.1
DPSeg (Inference I)	ConvNeXt-L	-	COCO-Stuff	<u>14.9</u>	<u>23.5</u>	<u>36.4</u>	<u>62.0</u>	<u>97.4</u>
DPSeg (Inference II, Ours)	ConvNeXt-L	-	COCO-Stuff	15.7	24.1	37.1	62.3	98.5

Table 1. Comparison with state-of-the-art methods on five open-vocabulary semantic segmentation test sets. mIoU results are shown, with the best in bold and the second-best underlined. In both configurations of the VLM model, our DPSeg under Inference I and II achieves superior performance, with Inference II consistently outperforming all other SOTA methods.

training images and $2k$ validation images and is split into two test sets: *A-150* with 150 annotated classes, and *A-847* with 847 common semantic categories.

Metrics. We evaluate our model using Intersection over Union (IoU). Our model is evaluated without resorting to finetuning in various settings, demonstrating the enhancements it achieves.

5.2. Implementation Details

Network Architectures. We employ the OpenAI pre-trained CLIP model [37] as the backbone, using either ConvNeXt-B or ConvNeXt-L as the image encoder and visual prompt encoder. For the text encoder, the text embedding dimension D_z is set to 640 for ConvNeXt-B and 768 for ConvNeXt-L. The initial cost volume fed into the decoder is set to a dimension of 128, with the output hidden dimensions of each decoder stage as [62, 32, 16]. For consistency, visual prompts are generated at 768×768 resolution, matching the input image size during training and inference. To improve training efficiency, the text and visual prompt encoders are frozen, with only the image encoder and cost volume-guided decoder being trained. Training uses a per-pixel binary cross-entropy loss. Our models are implemented in PyTorch [34] and Detectron2 [50]. We employ the AdamW optimizer with an initial learning rate of 2×10^{-4} , weight decay of 1×10^{-4} , and an additional factor $\alpha = 0.01$ to mitigate overfitting. The mini-batch size is set

to 4, and models are trained for $80k$ iterations on two V100 GPUs.

5.3. Quantitative Results

As shown in Table 1, in the base-VLM configuration, our approach achieves significant gains over state-of-the-art methods like SED [52], with improvements of +1.1%, +1.5%, +2.5%, and +1.1% mIoU on the *A-187*, *PC-459*, *A-150*, and *PC-59* datasets, respectively. Between our two inference stages, Inference II shows an additional gain of +0.87% mIoU over Inference I. In the *Large*-scale VLM configuration, our method further outperforms SED, achieving gains of +1.8%, +1.5%, +1.9%, +1.7%, and +2.4% mIoU across all datasets, with Inference II yielding an extra +0.92% mIoU over Inference I. These results underscore the effectiveness of our proposed method in advancing OVSS.

5.4. Qualitative Results

Fig. 6 shows segmentation examples from our model. In the first four columns, we display samples from the ADE-150 and PC-59 datasets alongside the ground truth, illustrating our model’s strong performance across diverse scenarios. Our segmentation results not only capture large foreground objects in both seen and unseen classes but also detect additional categories not annotated in the ground truth, such as “wood,” “fence,” and “chair” in the first, second,



Figure 6. Visualization examples of our model on open vocabulary semantic segmentation. **First four columns:** We present results from our model alongside the corresponding ground truth. **Fifth and sixth columns:** SED [52] results are shown for comparison, highlighting our model’s improved segmentation of small foreground objects, such as the mouse, bottle, and tennis racket.

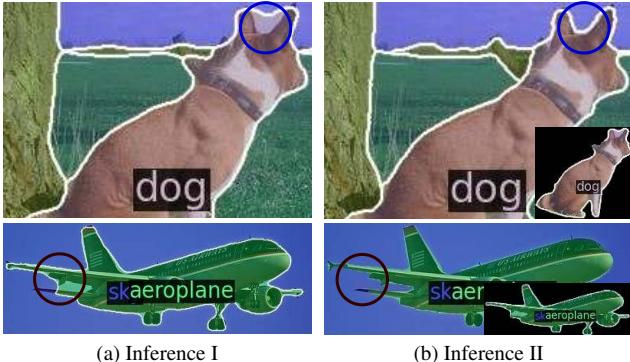


Figure 7. Comparison between inference I and II with our refinement strategy.

and fourth images of the first row. We also compare three cases between SED [52] and our model. As shown in the last two columns of Fig. 6, our method accurately segments smaller objects near larger ones, such as the mouse, bottle, and tennis racket, effectively preserving semantic information across scales. In Fig. 7, we compare two cases of our model between Inference I and Inference II. The results demonstrate our method’s effectiveness in enhancing semantic alignment and capturing finer segmentation details based on the initial segmentation output.

5.5. Ablation Study

For simplicity, all ablation studies are conducted using the strategy and results from Inference I.

Prompt strategy and cost volume generation. In this section, we conduct ablation studies to explore how different prompt strategies and cost volume generation strategies affect the segmentation results. From Table 2 (a), we observe that the visual-only prompt outperforms the text-only prompt. By averaging the visual and text embedding, achieving better overall performance. In Table 2 (b), we explore two additional cost volume generation methods: ① generating separate cost volumes for text and visual prompts, then concatenating them, and ② directly averaging the two cost volumes. The results clearly show that our approach—averaging the embeddings of both modalities before calculating cosine similarity—yields superior performance.

Cost volume-guided decoder. We examine the effect of various cost volume guidance strategies in our decoder, as shown in Table 3. When using only the cost volume \mathcal{F}_c generated by text and visual embeddings, segmentation performance is lowest (first row). Performance steadily improves as we progressively integrate multi-scale visual cost volumes $\mathcal{F}_v^{2,3,4}$, achieving an average increase of 1.8% mIoU

	Prompt Strategy	A-847	PC-459	A-150	PC-59	PAS-20
(a)	T	10.4	17.4	30.6	56.2	93.4
	V	11.1	18.0	31.2	56.9	94.5
	Avg(T, V) (Ours)	12.0	19.5	32.9	58.1	96.0
(b)	Cost Volume Generation	A-847	PC-459	A-150	PC-59	PAS-20
	Cat($\cos(\mathbf{T}, \mathbf{E}), \cos(\mathbf{V}, \mathbf{E})$)	11.7	19.2	32.3	57.8	95.4
	Avg($\cos(\mathbf{T}, \mathbf{E}), \cos(\mathbf{V}, \mathbf{E})$)	11.9	19.4	32.7	58.0	95.7
	\mathcal{F}_c (Ours)	12.0	19.5	32.9	58.1	96.0

Table 2. **Ablation study on prompt types and cost volume generation strategies.** Results are shown for text-only (**T**), visual-only (**V**), and combined text-visual prompt embeddings, along with comparisons of three cost volume generation solutions

Fusion Strategy	A-847	PC-459	A-150	PC-59	PAS-20
\mathcal{F}_c	10.6	17.9	31.6	57.0	94.9
$\mathcal{F}_c, \mathcal{F}_v^2$	11.4	18.8	32.1	57.6	95.5
$\mathcal{F}_c, \mathcal{F}_v^{2,3}$	11.7	19.2	32.5	57.9	95.7
$\mathcal{F}_c, \mathcal{F}_v^{2,3,4}$ (Ours)	12.0	19.5	32.9	58.1	96.0
$\mathcal{F}_c, \mathcal{F}_c(2\times, 4\times, 8\times)$	11.4	19.0	32.4	58.0	95.6

Table 3. **Ablation study on cost volume guidance strategies.** \mathcal{F}_c denotes the cost volume in our model, while \mathcal{F}_v^i represents cost volume at layer i , calculated from different levels of the image features and visual prompts. $\mathcal{F}_c(2\times)$ refers to the method where the cost volume \mathcal{F}_c is upsampled and concatenated.

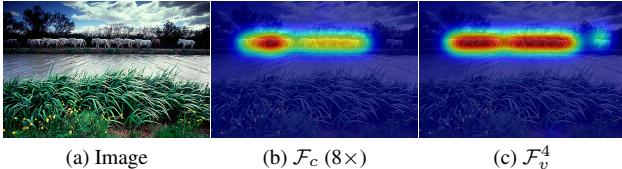


Figure 8. **Comparison of the visualized cost volumes.** Evaluating the effects of cost volume by comparing the (b) $8\times$ upsampled cost volume with the (c) the cost volume derived from 4th intermediate feature and visual prompt.

across all datasets when all scales are incorporated (second and third rows). Replacing $\mathcal{F}_v^{2,3,4}$ with an upsampled \mathcal{F}_c results in a performance drop (last row). Fig. 8 illustrates how our multi-scale cost volume captures finer semantic details, such as a single horse standing apart from a crowd—details that upsampled cost volumes fail to capture. This progressive integration of multi-scale cost volumes preserves spatial details, enhances multi-scale feature utilization, and maintains semantic consistency across scales, enabling our model to capture fine details for improving segmentation accuracy.

Number of Templates. Table 4 shows the results of an ablation study examining the impact of the number of templates on segmentation performance. In the experiment, each text template is paired with a corresponding visual prompt prototype. As the number of templates increases,

Number of Templates	A-847	PC-459	A-150	PC-59	PAS-20
1	3.6	7.8	15.6	32.1	72.3
10	6.9	11.2	20.3	41.3	81.7
20	7.8	14.9	27.9	47.3	86.7
40	10.1	16.8	29.6	55.9	91.8
80	12.0	19.5	32.9	58.1	96.0

Table 4. **Ablation study on the number of prompt templates.** Both text and visual prompts are used.

Visual Prompt Generation	A-847	PC-459	A-150	PC-59	PAS-20
(1) 20% Gaussian Noise	12.0	19.2	32.5	57.9	95.5
(2) 60% Gaussian Noise	11.6	18.8	32.9	57.6	95.0
(3) 80% Gaussian Noise	11.5	18.7	31.7	57.3	94.7
(4) LayerDiffuse[62]	12.1	19.4	32.7	58.0	95.8
(5) SD-V1.5	12.0	19.5	32.9	58.1	96.0
(6) SD-V3[10]	12.2	19.6	32.8	58.3	96.3
(7) SED (text prompt only)	11.4	18.6	31.6	57.3	94.4

Table 5. **Ablation study on visual prompts generation.** Methods(1)–(3) use SD-v1.5 with varying Gaussian noise levels (e.g., 80% means noise variance is 80% of the maximum), methods(4)–(6) employ LayerDiffuse, SD-v1.5, and SD-v3, respectively, and method(7) shows SED segmentation performance.

we observe a significant improvement in performance. This result suggests that a diverse set of templates better captures variations in real-world scenarios, as a single prompt may fail to represent all contexts. For example, ‘a photo of a bird’ may not accurately depict an image with multiple birds, which would be better represented by ‘a photo of many birds’.

Quality of visual prompts. The quality of visual prompts generated by different generation models varies, as shown in Table 5. Our experimental results highlight the robustness of our approach: even when using low-quality visual prompts (e.g., 80% Gaussian noise), our method consistently outperforms the text-prompt-only baseline (e.g., SED) in segmentation. Furthermore, employing different text-to-image generation models ((4)–(6) in Table 5) leads to minimal performance variation. These findings reinforce our commitment to integrating advanced generation techniques to enhance OVSS for both research and real-world applications.

6. Conclusion

In conclusion, we introduce **DPSeg**, a novel dual-prompt framework for open-vocabulary semantic segmentation (OVSS). Our approach synergistically combines text and visual prompts with multi-scale features, while introducing an innovative refinement strategy to enhance semantic alignment and fine-grained detail preservation. Extensive benchmarks validate its state-of-the-art performance, demonstrating the effectiveness of our dual-prompt cost volume learning approach.

References

- [1] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1348–1357, 2023.
- [2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.
- [3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.
- [4] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016.
- [5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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.
- [6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.
- [7] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.
- [8] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022.
- [9] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022.
- [10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- [11] Mark Everingham and John Winn. The pascal visual object classes challenge 2011 (voc2011) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 8, 2011.
- [12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.
- [13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.
- [14] Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Ji-ajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, et al. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 797–807, 2023.
- [15] Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. Clip-s4: Language-guided self-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11207–11216, 2023.
- [16] Mir Rayat Imtiaz Hossain, Mennatullah Siam, Leonid Sigal, and James J Little. Visual prompting for generalized few-shot segmentation: A multi-scale approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23470–23480, 2024.
- [17] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022.
- [18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.
- [19] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7661–7671, 2023.
- [20] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019.
- [21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.
- [22] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):453–465, 2013.
- [23] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.
- [24] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

- [25] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [26] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. *arXiv preprint arXiv:2312.04089*, 2023.
- [27] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024.
- [28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.
- [30] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.
- [31] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.
- [32] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*, pages 529–544. Springer, 2022.
- [33] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023.
- [34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.
- [35] Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, and Jiaya Jia. Hierarchical dense correlation distillation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23641–23651, 2023.
- [36] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020.
- [37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.
- [40] Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, and Changxin Gao. Open-vocabulary semantic segmentation with image embedding balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28412–28421, 2024.
- [41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.
- [42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.
- [45] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- [46] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 9197–9206, 2019.
- [47] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, and Liu Ren. USE: Universal Segment Embeddings for Open-Vocabulary Image Segmentation . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4187–4196, Los Alamitos, CA, USA, 2024. IEEE Computer Society.
- [48] Xiaoqi Wang, Wenbin He, Xiwei Xuan, Clint Sebastian, Jorge Piazentin Ono, Xin Li, Sima Behpour, Thang Doan, Liang Gou, Han-Wei Shen, et al. Use: Universal segment embeddings for open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4187–4196, 2024.

- [49] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [50] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. <https://github.com/facebookresearch/detectron2>, 2019.
- [51] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019.
- [52] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3426–3436, 2024.
- [53] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022.
- [54] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023.
- [55] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022.
- [56] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023.
- [57] Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2771–2781, 2023.
- [58] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.
- [59] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.
- [60] Canyu Zhang, Xiaoguang Li, Qing Guo, and Song Wang. Sair: Learning semantic-aware implicit representation. In *European Conference on Computer Vision*, pages 319–335. Springer, 2024.
- [61] Gengwei Zhang, Guoliang Kang, Yi Yang, and Yunchao Wei. Few-shot segmentation via cycle-consistent trans-
- former. *Advances in Neural Information Processing Systems*, 34:21984–21996, 2021.
- [62] Lvmin Zhang and Maneesh Agrawala. Transparent image layer diffusion using latent transparency, 2024.
- [63] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.
- [64] Ziyu Zhao, Zhenyao Wu, Xinyi Wu, Canyu Zhang, and Song Wang. Crossmodal few-shot 3d point cloud semantic segmentation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4760–4768, 2022.
- [65] Ziyu Zhao, Pingping Cai, Canyu Zhang, Xiaoguang Li, and Song Wang. Crossmodal few-shot 3d point cloud semantic segmentation via view synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8777–8785, 2024.
- [66] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luwei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.
- [67] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.
- [68] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.
- [69] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023.