# Images as Noisy Labels: Unleashing the Potential of the Diffusion Model for Open-Vocabulary Semantic Segmentation

Anonymous ICCV submission

Paper ID 8878

## Abstract

*Recently, open-vocabulary semantic segmentation has garnered growing attention. Most current methods leverage vision-language models like CLIP to recognize unseen categories through their zero-shot capabilities. However, CLIP struggles to establish potential spatial dependencies among scene objects due to its holistic pre-training objective, causing sub-optimal results. In this paper, we propose a DEnoising learning framework based on the Diffusion model for Open-vocabulary semantic Segmentation, called DEDOS, which is aimed at constructing the scene skeleton. Motivation stems from the fact that diffusion models incorporate not only the visual appearance of objects but also embed rich scene spatial priors. Our core idea is to view images as labels embedded with "noise"—non-essential details for perceptual tasks—and to disentangle the intrinsic scene prior from the diffusion feature during the denoising process of the images. Specifically, to fully harness the scene prior knowledge of the diffusion model, we introduce learnable proxy queries during the denoising process. Meanwhile, we leverage the robustness of CLIP features to texture shifts as supervision, guiding proxy queries to focus on constructing the scene skeleton and avoiding interference from texture information in the diffusion feature space. Finally, we enhance spatial understanding within CLIP features using proxy queries, which also serve as an interface for multi-level interaction between text and visual modalities. Extensive experiments validate the effectiveness of our method, experimental results on five standard benchmarks have shown that DEDOS achieves state-of-the-art performance. We will make the code publicly available.*

## 1. Introduction

Image segmentation is a core task in computer vision, aimed at assigning semantic labels to pixels in an image. Despite achieving excellent performance in recent years [6, 9, 10, 15, 27, 51], traditional methods are often designed for predefined sets of training categories, leading to limitations in handling diverse visual concepts. To overcome this challenge, open vocabulary semantic segmentation [13, 16, 25, 47, 53] has emerged as an approach in which segmentation models are trained to recognize objects of arbitrary categories based on the given text, broadening their applicability in real-world scenarios.

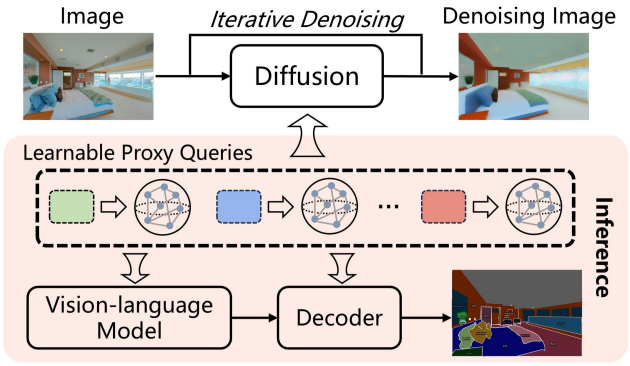


Figure 1. Our method treats images as labels embedded with noise—irrelevant details to perceptual tasks—and employs learnable proxy queries to progressively construct the scene skeleton from diffusion features during the iterative denoising. These proxy queries are ultimately used to improve the spatial understanding of the vision-language model, avoiding visual shortcomings in dense perception tasks that arise from its holistic pre-training objective.

Most current approaches for open-vocabulary semantic segmentation [16, 47–49, 53, 54] leverage the powerful zero-shot ability of large-scale vision-language models, such as CLIP [33] and ALIGN [19], which are pre-trained on extensive image-text datasets. While these pre-trained models excel at classifying single object proposals or images, their image-level pre-training objective introduces ambiguity in spatial relationships [40, 49] and confusion with co-occurring objects, leading to severe visual shortcomings in dense perception tasks.

On the other hand, diffusion models [12, 35], i.e., generative models based on denoising, have made great breakthroughs in generating diverse and high-fidelity im-

ages. Given the powerful generative capabilities of diffusion models, harnessing them to construct generation-based perceptual frameworks opens new avenues for advancing perceptual tasks, such as object detection [7], semantic segmentation [1, 8, 45], and depth estimation [18]. Unlike previous approaches that simply use diffusion models as a data generator [32, 42, 43] or an alternative feature extractor [1, 25], we focus on exploring how to effectively leverage the diffusion model's prior knowledge for enhanced scene understanding. This prior knowledge is reflected in the fact that with only a few keywords, such as cars, people, and roads, the diffusion model can automatically infer distributional associations between different objects in the scene to produce high-fidelity images.

Therefore, this paper seeks to unleash the potential of diffusion models for open-vocabulary semantic segmentation by parsing their intrinsic scene priors to construct the scene skeleton—i.e., the spatial dependencies or contextual relationships among objects—which allows the model to transcend individual object recognition, establish rich semantic relationships, and achieve a deeper understanding of the spatial arrangement within the scene. The intuition behind our work is the following: (1) *If a comprehensive and encyclopedic representation of the real world indeed underpins open-vocabulary semantic segmentation, then models capable of segmenting any text-specified category can be derived from pre-trained diffusion models.* (2) *Any model with strong zero-shot generalization must possess a complete understanding of the scene skeleton.*

However, the direct use of pre-trained diffusion models for perceptual tasks is non-trivial because they are primarily designed to synthesize images through iterative denoising rather than to provide strong representations. In this paper, we propose a denoising learning framework based on the diffusion model for the open-vocabulary semantic segmentation, dubbed as DEDOS. Our insight is to view the image as a label embedded with noise and to construct the scene skeleton from diffusion features during the denoising process of the images, as illustrated in Figure 1. Specifically, we employ learnable proxy queries to progressively construct the scene skeleton throughout the diffusion model's iterative denoising process. Moreover, as the diffusion model is primarily built for generative tasks where feature space encompasses both scene prior knowledge and visual texture details, decoupling of diffusion features is necessary. To achieve this, we use CLIP features, which are robust to shifts in texture distribution, to guide proxy queries, effectively filtering out texture information that is irrelevant or detrimental to the perceptual task. Finally, we leverage proxy queries to refine the scene understanding within CLIP features, which also serve as an interface to facilitate multi-level interactions between the text and visual modalities. Across various benchmark datasets, DEDOS achieves state-of-the-art performance, surpassing existing methods by a considerable margin. Moreover, even in challenging scenarios where the destination category is drastically different from the training dataset, our model consistently surpasses current state-of-the-art methods by a substantial margin. We summarize our contribution as follows:

- To the best of our knowledge, DEDOS is the first denoising learning framework based on the diffusion model that views images as noisy labels and learns to construct scene skeletons throughout the denoising process.
- We employ proxy queries to progressively parse scene priors encoded in diffusion features, while leveraging CLIP's robustness to disentangle texture information and prevent disruption to proxy queries.
- Our framework, DEDOS, achieves state-of-the-art performance on standard benchmarks and in challenging extreme-case scenarios, showcasing its versatility and practicality.

## 2. Related Works

### 2.1. Diffusion Model

Diffusion models [12, 17, 35, 39] have achieved remarkable success on visual generation tasks such as text-to-image synthesis [34–36], inpainting [29], and image editing [23, 30]. Recent research has investigated the application of diffusion models for various perception tasks, including semantic segmentation [1, 45], object detection [7] and depth estimation [18, 24]. For example, Jia et al. [20] introduce a pipeline that leverages the rich prior knowledge of a pre-trained diffusion model to generate task-specific images with diverse visual characteristics. Xu et al. [45] propose ODISE, which integrates pre-trained text-to-image diffusion models to perform open-vocabulary panoptic segmentation. Chen et al. [7] introduce DiffusionDet, a framework that reinterprets object detection as a generative process, refining noisy bounding boxes into accurate object boxes. Benigmim et al. [2] employ diffusion models to enhance source image diversity, aiming to capture various target domain patterns. Unlike the above methods that mainly treat diffusion models as another form of feature extractor or data generator, we investigate how to leverage the rich prior knowledge of the scenarios in diffusion models to enhance the zero-shot segmentation capability of the models.

### 2.2. Open-vocabulary Semantic Segmentation

Open-vocabulary semantic segmentation seeks to divide an image into semantic regions based on arbitrary textual descriptions. Most current methods, building on the success of large-scale vision-language models in zero-shot classification [19, 33], explore ways to adapt multimodal image-text alignment to a finer level of granularity [16, 25, 47–49, 53, 54]. OpenSeg [16] divides the task into a region pro-

poser that generates mask regions and a grounder that links them to words in captions. MaskCLIP [53] refines generated proposals by utilizing the attention maps from a frozen CLIP model. On the other hand, ZegCLIP [54] presents a single-stage framework that leverages CLIP embeddings for direct mask prediction. SAN [47] integrates a side network with CLIP to generate regions and determine their corresponding semantic categories. CAT-Seg [11] introduces a cost aggregation framework that leverages cosine-similarity scores between images and text to fine-tune vision-language models. Previous approaches have aimed at refining the feature granularity of large-scale vision-language models to adapt image-level pre-trained models for pixel-level tasks. In contrast, we address the spatial perception limitations of vision-language models by leveraging the scene prior knowledge embedded in diffusion models, facilitating the construction of the scene skeleton.

### 2.3. Learnable Query Design

Motivated by DETR [5], several frameworks built upon learnable queries have emerged recently. Mask2Former [10] leverages object queries to cluster pixels, unifying semantic, instance, and panoptic segmentation tasks into a single framework. ECENet [26] derives object queries from predicted class masks and enhances them through explicit interactions with multi-stage image features. MDETR [21] is an end-to-end detection framework that leverages raw text queries and a transformer-based architecture for object detection, demonstrating strong performance in cross-modal tasks. Rein [41] leverages a set of learnable queries to adaptively fine-tune various vision foundation models, leading to substantial performance gains in cross-domain segmentation tasks. Although previous studies have demonstrated that learnable queries can effectively enhance the performance of perceptual tasks, how to use learnable queries to mine scenario prior information from diffusion models remains unexplored. Our work aims to establish a concise and efficient framework that employs learnable proxy queries to refine the scene comprehension of vision-language models.

## 3. Methodology

### 3.1. Problem Definition

Open-vocabulary semantic segmentation assigns a semantic label to each pixel in an image based on a set of categories defined by free-form text. During training, the model is only provided with a set of images and their annotations on base classes $\mathcal{C}_B$. In the inference stage, we evaluate the model on another set of images with a new set of novel classes $\mathcal{C}_N$, which are not encountered during training, i.e., $\mathcal{C}_B \neq \mathcal{C}_N$.

### 3.2. Preliminary of Stable Diffusion

Our research starts with stable diffusion (SD), which involves a diffusion process and a denoising process. In the diffusion process, a given image $x$ is first mapped to the latent space using a pre-trained encoder $\mathcal{V}$:

$$z_0 = \mathcal{V}(x) \tag{1}$$

where $z_0$ represents the latent code. Next, Gaussian noise is added to the latent code $z_0$ based on a pre-defined timestep $t$: $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{(1-\bar{\alpha}_t)}\epsilon$, where $\epsilon \sim \mathcal{N}(0,1)$ and $\bar{\alpha}_t$ represents the pre-defined noise scaling factors. The denoising process progressively converts noisy samples back into the original data. A single denoising step is defined as:

$$p_\theta(z_{t-1} \mid z_t) := \mathcal{N}(z_{t-1} \mid \mu_\theta(z_t, t), \Sigma_\theta(z_t, t)) \tag{2}$$

where $\mu_\theta$ is the predicted mean obtained from the noise predictor $\epsilon_\theta(\cdot)$, while $\Sigma_\theta$ is the pre-defined covariance. In the denoising process, $\epsilon_\theta(\cdot)$ is applied iteratively until it reaches the raw data $z_0$. The latent code can be converted to pixel space by a pre-trained decoder $\mathcal{R}$. In this paper, we treat the latent code corresponding to the original image $x$ as noisy samples $z_t$.

### 3.3. Overview

As shown in Figure 2, we present DEDOS, a denoising learning framework based on diffusion models for open-vocabulary semantic segmentation. DEDOS introduces learnable proxy queries to progressively construct the scene skeleton during the denoising process of the diffusion model (Sec 3.4). Given that the diffusion feature space encompasses both scene distribution priors and fine-grained texture details, we leverage CLIP's robustness to texture shifts to effectively disentangle the diffusion feature space (Sec 3.5). Finally, we strengthen scene perception of CLIP features by integrating optimized proxy queries, which also act as an interface for multi-level interactions between text and visual modalities (Sec 3.6). The details are as follows.

### 3.4. Proxy Queries for Scene Skeleton

In this section, by treating images as labels embedded with noise—non-essential textural details for perception—we expect prior knowledge of the scene skeleton, i.e., the implicit spatial dependencies between objects, to emerge during the iterative denoising process of the diffusion model.

**Excavating scene priors in diffusion models.** The key to effectively utilizing the pre-training prior of the diffusion model is to preserve its latent space. To achieve this, we introduce learnable proxy queries within the denoising process. This minimal disruption to the pre-trained diffusion model allows us to fully unleash its potentially powerful capabilities for scene understanding. Specifically, we employ learnable proxy queries to actively interact with features at each layer of the diffusion model. Precisely, learnable proxy queries are linearly projected into the queries $\mathbf{Q}$ and the features $f_i^{sd}$ produced by the $i$-th layer of the noise predictor

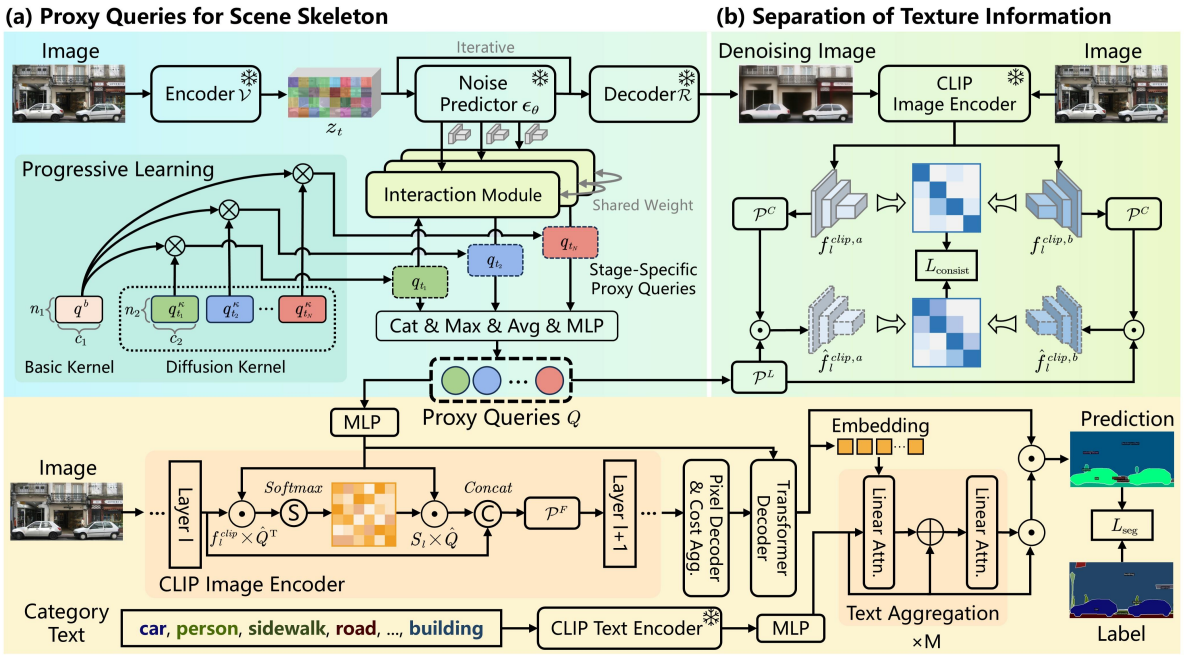


Figure 2. A brief illustration of our proposed framework. Our method introduces learnable proxy queries to progressively construct the scene skeleton during the denoising process of the diffusion model. Next, we leverage the robustness of CLIP's visual features to texture shifts to disentangle the diffusion feature space. Finally, we enhance the scene perception within CLIP features using optimized proxy queries, which also serve as an interface to facilitate multi-level interactions between textual and visual modalities.

$\epsilon_\theta(\cdot)$ are respectively projected into the keys $\mathbf{K}$ and values $\mathbf{V}$:

$$\mathbf{Q}_i = q_{i-1,t}\mathbf{W}_i^{\mathcal{Q}}, \quad \mathbf{K}_i = f_i^{sd}\mathbf{W}_i^{\mathcal{K}}, \quad \mathbf{V}_i = f_i^{sd}\mathbf{W}_i^{\mathcal{V}}. \quad (3)$$

where $\mathbf{W}_i^{\mathcal{Q}}$,$\mathbf{W}_i^{\mathcal{K}}$,$\mathbf{W}_i^{\mathcal{V}}$ are linear projections, and $t$ is the timestep of the diffusion model. The modifications of the proxy queries are calculated as:

$$\hat{q}_{i,t} = \text{softmax}\left(\mathbf{Q}_i\mathbf{K}_i^{\mathrm{T}}\mathcal{M}_i\right)\mathbf{V}_i\mathcal{M}_i + q_{i-1,t} \quad (4)$$

where the $\mathcal{M}_i$ is the mask map for the $i$-th layer of the diffusion model, which aims to facilitate the proxy queries to focus on the general perception of the scene distribution rather than single object proposals. To enhance the flexibility in feature adjustment, we utilize the MLP composed of $W_i$ and $b_i$ to produce the new proxy queries:

$$q_{i,t} = \hat{q}_{i,t} \times W_i + b_i \quad (5)$$

where $q_{i,t}$ is continuously used to interact with the next-layer diffusion features.

**Progressive learning.** Just as light is refracted through a prism, we encourage proxy queries to capture various levels of the scene skeleton, facilitating a gradual unfolding of the scene's underlying distribution. Specifically, we decompose the proxy queries into two parts: the basic kernel and the diffusion kernel. The basic kernel, shared across all denoising processes, provides a consistent foundation, while each diffusion kernel, unique to a specific denoising process, captures its distinctive characteristics. At the beginning of each denoising process $t$, stage-specific proxy queries $q_t$ are dynamically generated by applying the Kronecker product between the basic kernel $q^b$ and the corresponding diffusion kernel $q_t^\kappa$:

$$\begin{aligned} q_t &= q^b \otimes q_t^\kappa \\ q^b \in \mathbb{R}^{n_1\times c_1}, &\quad q_t^\kappa \in \mathbb{R}^{n_2\times c_2}, \quad q_t \in \mathbb{R}^{n\times c} \end{aligned} \quad (6)$$

where $n = n_1 n_2$ represents the number of queries, $c = c_1 c_2$ is the dimension of queries, and $\otimes$ is the Kronecker product matrix operation. Each $q_t$ corresponds to one denoising process, and at the beginning of the denoising process, $q_{0,t} = q_t$. At the end of the denoising process, the original $q_t$ is replaced by the resulting new $q_{L,t}$, where $L$ is the number of layers of the noise predictor $\epsilon_\theta(\cdot)$. Finally, we transform the diverse $q_t$ into a set of global proxy queries $Q$ by computing both the maximal component and the average component as follows:

$$Q_{avg} = \frac{1}{N}\sum_{j=1}^{N} q_{t_j}, \quad Q_{\max} = \max_{j=1,2,\ldots,N} q_{t_j} \quad (7)$$

where $N$ represents the number of iterative denoising steps. Subsequently, $Q$ is derived as:

$$Q = \text{Concat}\left([Q_{\max}, Q_{\text{avg}}, q_{t_1}, \ldots, q_{t_N}]\right) \times W_Q + b_Q \quad (8)$$

where $W_Q$ and $b_Q$ denote the weights and biases. Average queries suppress noise, while maximum queries highlight crucial features. Combined with stage-specific learnable queries, they enhance adaptability across diverse scenes.

### 3.5. Separation of Texture Information

Since the diffusion model is designed for generative tasks, its feature space encompasses not only rich scene priors but also visual texture information. Texture details, which may be encoded into proxy queries but are irrelevant or even harmful to open-vocabulary semantic segmentation, need to be disentangled from the proxy queries. Given that the primary differences between images before and after denoising lie in texture details, while scene layout remains nearly identical, we leverage CLIP embeddings as supervision due to their robustness to texture shifts, effectively guiding the disentanglement of information within proxy queries. Specifically, we feed the images before and after denoising into CLIP image encoder to extract their respective features:

$$f_l^{clip,b} = \text{CLIP}(x_b), \quad f_l^{clip,a} = \text{CLIP}(x_a) \tag{9}$$

where $l$ is the index of the CLIP layer, $x_b$ represents the image before denoising and $x_a$ represents the image after denoising. We generate new features $\hat{f}_l^{clip,b}$ and $\hat{f}_l^{clip,a}$ by interacting CLIP feature with the proxy queries $Q$:

$$\begin{aligned} \hat{f}_l^{clip,b} &= \mathcal{P}^L(Q) \times \mathcal{P}^C(f_l^{clip,b}) \\ \hat{f}_l^{clip,a} &= \mathcal{P}^L(Q) \times \mathcal{P}^C(f_l^{clip,a}) \end{aligned} \tag{10}$$

where $\mathcal{P}^L$ is the linear projection layer and $\mathcal{P}^C$ represents $1\times1$ convolution. We calculate the similarity matrices of the CLIP features before and after denoising, as well as the similarity matrices of the interaction features before and after denoising:

$$M_{clip} = \frac{f_l^{clip,b} \cdot f_l^{clip,a}}{\left\| f_l^{clip,b} \right\| \left\| f_l^{clip,a} \right\|}, M_{query} = \frac{\hat{f}_l^{clip,b} \cdot \hat{f}_l^{clip,a}}{\left\| \hat{f}_l^{clip,b} \right\| \left\| \hat{f}_l^{clip,a} \right\|} \tag{11}$$

Finally, we compute the consistency loss:

$$L_{\text{consist}} = L_\delta(M_{query}, M_{clip}) \tag{12}$$

where $L_\delta$ is Smooth $L_1$ Loss. This consistency constraint ensures that proxy queries prioritize scene layout over texture details by aligning them with CLIP's robust feature representations.

### 3.6. Mask Generation and Classification

**Feature refinement.** We use proxy queries $Q$ to enhance the scene perception within CLIP feature representation. Specifically, we employ a dot-product operation to generate a similarity map $S_l$:

$$S_l = \text{softmax}(f_l^{clip} \times \hat{Q}^{\text{T}}) \tag{13}$$

where $f_l^{clip}$ is the feature from the $l$-th layer of the CLIP image encoder, $\hat{Q} = \text{MLP}(Q)$ and MLP parameters are shared across layers. Using the similarity map $S_l$, we preliminarily estimate the new feature $\bar{f}_l = S_l \times \hat{Q}$. Finally, $\bar{f}_l$ and $f_l^{clip}$ are concatenated and fused via $1\times1$ convolution:

$$\hat{f}_l^{clip} = \mathcal{P}^F(\text{Concat}([\bar{f}_l, f_l^{clip}])) \tag{14}$$

where $\mathcal{P}^F$ is $1\times1$ convolution, $\hat{f}_l^{clip}$ will be fed to the next layer of the CLIP image encoder.

**Cross-modal interaction.** We use proxy queries $Q$ as an interface to facilitate multi-level interactions between textual and visual modalities. To achieve this, we propose adding a text aggregation module to the decoder. A typical decoder consists of a pixel decoder, which extracts features, and a transformer decoder, which outputs the predicted mask [10]. We insert the text aggregation module after the transformer decoder, which consists of linear transformer blocks [22]. Specifically, CLIP features are fed to the pixel decoder to get multi-scale features. Proxy queries $Q$ then perform masked cross-attention with these features in the transformer decoder to obtain both the predicted mask and object embeddings $Q_{object}$. The text aggregation module integrates $Q_{object}$ with the text embeddings to produce new object embeddings $\hat{Q}_{object}$. Finally, classification information is obtained by computing the dot product between $\hat{Q}_{object}$ and text embeddings, followed by a sigmoid activation to produce class probabilities. The segmentation loss $L_{\text{seg}}$ is defined as follows:

$$L_{\text{seg}} = \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{bce}} L_{\text{bce}} + \lambda_{\text{dice}} L_{\text{dice}} \tag{15}$$

where $L_{\text{cls}}$ represents the cross-entropy loss for classification, while $L_{\text{bce}}$ and $L_{\text{dice}}$ correspond to the binary cross-entropy and dice loss for the predicted mask, respectively. The loss weights $\lambda_{\text{bce}}$, $\lambda_{\text{dice}}$ and $\lambda_{\text{cls}}$ are set equally.

**Training objective.** The overall training objective comprises the segmentation loss and the consistency loss:

$$L_{\text{total}} = L_{\text{seg}} + \gamma L_{\text{consist}} \tag{16}$$

where $\gamma$ is a hyperparameter. Notably, the inference stage omits steps related to the diffusion model and the separation of texture information, ensuring that our method remains concise, efficient, and broadly applicable.

## 4. Experiments

### 4.1. Datasets and Evaluation

We train our model on the COCO-Stuff [4] with 118k annotated images across 171 categories and evaluate it on both standard open-vocabulary semantic segmentation benchmarks and a multi-domain dataset with domain-specific images. We evaluate performance using the mean Intersection-over-Union (mIoU), following standard practice.

| Model | VLM | Training Dataset | Additional Dataset | PC-459 | A-150 | PC-59 | PAS-20 | PAS-$20^b$ |
|---|---|---|---|---|---|---|---|---|
| OpenSeg [16] | ALIGN | COCO Panoptic | ✓ | 7.9 | 17.5 | 40.1 | - | 63.8 |
| OVSeg [25] | CLIP ViT-B/16 | COCO-Stuff | ✓ | 11.0 | 24.8 | 53.3 | 92.6 | - |
| ZegCLIP [54] | CLIP ViT-B/16 | COCO-Stuff-156 | ✗ | - | - | 41.2 | 93.6 | - |
| SAN [47] | CLIP ViT-B/16 | COCO-Stuff | ✗ | 12.6 | 27.5 | 53.8 | 94.0 | - |
| EBSeg [38] | CLIP ViT-B/16 | COCO-Stuff | ✗ | 17.3 | 30.0 | 56.7 | 94.6 | - |
| SED [44] | ConvNeXt-B | COCO-Stuff | ✗ | 18.6 | 31.6 | 57.3 | 94.4 | - |
| CAT-Seg [11] | CLIP ViT-B/16 | COCO-Stuff | ✗ | 19.0 | 31.8 | 57.5 | 94.6 | 77.3 |
| DEDOS (Ours) | CLIP ViT-B/16 | COCO-Stuff | ✗ | **21.3** | **33.7** | **59.8** | **95.7** | **80.2** |
| OpenSeg [16] | ALIGN | COCO Panoptic | ✓ | 11.5 | 26.4 | 44.8 | - | 70.2 |
| OVSeg [25] | CLIP ViT-L/14 | COCO-Stuff | ✓ | 12.4 | 29.6 | 55.7 | 94.5 | - |
| SAN [47] | CLIP ViT-L/14 | COCO-Stuff | ✗ | 15.7 | 32.1 | 57.7 | 94.6 | - |
| ODISE [45] | CLIP ViT-L/14 | COCO-Stuff | ✗ | 14.5 | 29.9 | 57.3 | - | - |
| EBSeg [38] | CLIP ViT-L/14 | COCO-Stuff | ✗ | 21.0 | 32.8 | 60.2 | 96.4 | - |
| SED [44] | ConvNeXt-L | COCO-Stuff | ✗ | 22.6 | 35.2 | 60.6 | 96.1 | - |
| CAT-Seg [11] | CLIP ViT-L/14 | COCO-Stuff | ✗ | 23.8 | 37.9 | 63.3 | 97.0 | 82.5 |
| DEDOS (Ours) | CLIP ViT-L/14 | COCO-Stuff | ✗ | **25.6** | **39.4** | **65.7** | **97.6** | **84.6** |

Table 1. **Quantitative comparison with state-of-the-art methods on standard benchmarks.** A, PC, and PAS denote ADE20K [52], Pascal Context [31], and Pascal VOC [14], respectively. The best results are highlighted in bold.

| Model | VLM | General | Earth Monit. | Medical Sciences | Engineering | Agri. and Biology | Mean |
|---|---|---|---|---|---|---|---|
| *Random (LB)* | - | *1.2* | *7.1* | *29.5* | *11.7* | *6.1* | *10.3* |
| *Best supervised (UB)* | - | *48.6* | *79.1* | *89.5* | *67.7* | *81.9* | *71.0* |
| ZSSeg [46] | CLIP ViT-B/16 | 20.0 | 18.0 | 41.8 | 14.0 | 22.3 | 22.7 |
| ZegFormer [13] | CLIP ViT-B/16 | 13.6 | 17.3 | 17.5 | 17.9 | 25.8 | 17.6 |
| X-Decoder [55] | UniCL-T | 22.0 | 18.9 | 23.3 | 15.3 | 18.2 | 19.8 |
| OpenSeeD [50] | UniCL-B | 22.5 | 25.1 | **44.4** | 16.5 | 10.4 | 24.3 |
| SAN [47] | CLIP ViT-B/16 | 29.4 | 30.6 | 29.9 | **23.6** | 15.1 | 26.7 |
| CAT-Seg [11] | CLIP ViT-B/16 | 38.7 | 35.9 | 28.1 | 20.3 | 32.6 | 32.0 |
| DEDOS (Ours) | CLIP ViT-B/16 | **42.6** | **38.1** | 41.3 | 22.6 | **34.7** | **35.9** |
| OVSeg [25] | CLIP ViT-L/14 | 29.5 | 29.0 | 31.9 | 14.2 | 28.6 | 26.9 |
| SAN [47] | CLIP ViT-L/14 | 36.2 | 38.8 | 30.3 | 17.0 | 20.4 | 30.1 |
| CAT-Seg [11] | CLIP ViT-L/14 | 44.7 | 40.0 | 24.7 | 20.2 | 38.6 | 34.7 |
| DEDOS (Ours) | CLIP ViT-L/14 | **46.9** | **41.8** | **40.6** | **23.5** | **41.2** | **38.8** |

Table 2. **Quantitative comparison with state-of-the-art methods on MESS [3].** MESS covers diverse domain-specific datasets, which present significant challenges due to their differences from the training dataset. We present the average score for each domain. See supplementary material for detailed results. *Random* denotes the lower bound from uniform distributed prediction, while *Best supervised* represents the upper bound for dataset performance. The best results are highlighted in bold.

**Datasets for standard benchmarks.** For standard benchmarks, we evaluate our model on ADE20K [52], PASCAL VOC [14], and PASCAL-Context [31] datasets following previous works [11, 45, 47]. ADE20K has 2k validation images with 150 common classes (A-150). PASCAL-Context contains 5k images and 459 classes (PC-459), or 59 frequent classes (PC-59). PASCAL VOC includes 20 object classes and a background class, with 1.5k images for validation. We report the results for PAS-20 (containing 20 object classes) and PAS-$20^b$ (containing 20 object classes and a background class), as in Cho et al. [11].

**Datasets for multi-domain evaluation.** We further evaluate our model on the MESS benchmark [3], which includes 22 datasets to assess the real-world effectiveness of open-vocabulary models. These datasets cover a broad range of fields like earth monitoring, engineering, agriculture, and general domains such as driving and paintings. We present the average scores for each domain in the main text. Detailed results for all 22 datasets can be found in the supplementary material.

### 4.2. Implementation Details

We utilize the Detectron2 codebase for our implementation. We adopt CLIP [33] as the backbone and use the Mask2Former [10] decoder, a widely-used segmentation


CAT-Seg Ours Ground truth CAT-Seg Ours Ground truth

Figure 3. Qualitative results on ADE20K validation sets. DEDOS demonstrates more accurate category predictions and more complete spatial distributions. The supplement contains more visual results.

head, further enhancing it with cost aggregation [11] applied after its pixel decoder. For the training phase, we use the AdamW optimizer [28], setting the learning rate at $2\times10^{-6}$ for the backbone and $2\times10^{-4}$ for both the decoder and the learnable queries. We train all models for 60k iterations with a batch size of 4, cropping images to a resolution of 640 × 640. Training is performed on 2 NVIDIA A100 GPUs. For the diffusion model, we utilize Stable Diffusion v2-1 [35], pretrained on LAION5B [37], and keep it frozen throughout training.

### 4.3. Comparison with State-of-the-art Methods

**Results of standard benchmarks.** Table 1 compares our method with other competing approaches on standard open-vocabulary semantic segmentation benchmarks, including ODISE [45], the previous SOTA diffusion-based method (refer to the supplementary material for further comparisons with diffusion-based methods). Overall, we significantly outperform all previous approaches in both scales of backbone setups, including ViT-B/16 and ViT-L/14. Even when compared to approaches that leverage additional datasets for performance gains, our method still significantly outperforms all competing approaches. In particular, when using ViT-B/16, our method significantly outperforms the previous approach CAT-Seg, which also uses ViT-B/16, even matching or exceeding the performance of methods using the more powerful ViT-L/14 backbone. When employing the ViT-L/14 model as the backbone, our method demonstrates remarkable results, achieving 25.6 mIoU in the challenging PC-459 dataset, clearly outperforming others. This emphasizes the effectiveness of our method in generalizing to a variety of real-world scenarios.

**Results of multi-domain evaluation.** Table 2 presents the quantitative results of our method on the MESS benchmark [3]. Generally, our model achieves the highest mean score, outperforming other models with a substantial performance boost, particularly in the general domain as well as in agriculture and biology. In the fields of medical science and engineering, where multispectral and electromagnetic images are prevalent, we speculate that diffusion models may possess limited understanding when dealing with images from these specialized spectra.

| | Components | PC-459 | A-150 | PC-59 | PAS-20 | PAS-$20^b$ |
|---|---|---|---|---|---|---|
| **(I)** | Baseline | 22.8 | 36.0 | 60.6 | 96.5 | 79.6 |
| **(II)** | **(I)** + Proxy queries (Once). | 23.8 | 37.2 | 62.9 | 97.0 | 81.9 |
| **(III)** | **(II)** + Progressive learning. | 24.7 | 38.6 | 64.2 | 97.2 | 83.1 |
| **(IV)** | **(III)** + Decoupling. | 25.4 | 39.1 | 64.9 | 97.5 | 84.0 |
| **(V)** | **(IV)** + Text interaction. | **25.6** | **39.4** | **65.7** | **97.6** | **84.6** |

Table 3. **Ablation study on the effect of main components.** We conduct the ablation study by gradually adding components to the baseline. "Once" signifies a single denoising step.

**Qualitative results.** We visually compare the segmentation results with the previous SOTA method CAT-Seg on ADE20K validation sets in Figure 3. Notably, our method produces more accurate category predictions, such as 'floor' in the first row on the left of Figure 3, as well as more complete spatial coverage, such as 'car' in the first row on the right of Figure 3. We attribute this improvement to our method's effective learning of the scene skeleton, which plays a critical role in open-vocabulary semantic segmentation.

### 4.4. Ablation Studies

In this section, we present comprehensive experiments to validate the effectiveness of our method. Please refer to supplementary material for more results and analysis.

**Effect of components.** We evaluate the effectiveness of the primary components on standard open-vocabulary semantic

| Model | +SD 1.4 | +SD 1.5 | +SD 2.1 |
|---|---|---|---|
| mIoU | 65.4 | 65.6 | 65.7 |

Table 4. Quantitative comparison of different stable diffusion models on the PC-59 dataset.

| Number of iterations | 2 | 3 | 5 | 7 |
|---|---|---|---|---|
| mIoU | 65.0 | 65.7 | 65.7 | 65.2 |

Table 5. Quantitative comparison of different iteration numbers on the PC-59 dataset.

| Timesteps | [50, 75, 150] | [50, 100, 200] | [50, 200, 300] | [100, 200, 300] |
|---|---|---|---|---|
| mIoU | 65.2 | 65.7 | 64.9 | 64.6 |

Table 6. Ablation results of different diffusion timesteps on the PC-59 dataset.

segmentation benchmarks. The baseline model builds on CLIP [33] as the backbone, integrating the Mask2Former decoder [10] and further enhancing it with cost aggregation [11] applied after its pixel decoder. As shown in Table 3, we first add the learnable proxy queries to the baseline in (II), which significantly improves performance. Next, we decompose the proxy query into the basic kernel and the diffusion kernel, enabling progressive learning in (III) that further improves performance compared to (II). Then, after decoupling the diffusion features, the segmentation results continue to improve in (IV). Finally, by establishing interaction with text embeddings, our model achieves optimal results on all datasets in (V), significantly outperforming previous approaches.

**Comparing different stable diffusion.** As shown in Table 4, we perform a comparative experiment with three currently dominant stable diffusion models. The results indicate that our approach delivers consistent performance across different diffusion models, demonstrating that the final outcomes are not significantly affected by the choice of diffusion model. This highlights the robustness and adaptability of our method.

**The choice of iteration numbers.** To gradually construct the scene skeleton, we utilize the diffusion model to iteratively denoise the image. As shown in Table 5, the model achieves the best performance when the number of iterations is set to 3 and 5. We ultimately choose 3 as the default parameter, which is applied in subsequent experiments.

**The choice of timesteps.** We further investigate the effect of choosing different timesteps on the model results. Timesteps $t$ correspond to distinct denoising stages, with a larger timestep $t$ corresponding to a higher weight of noise. As shown in Table 6, choosing different timesteps has a large effect on the results of the model. We empirically adapt $t = [50, 100, 200]$ as the default parameter.

**Study on the number and dimensions of proxy queries.** As shown in Figure 4, we explore proxy query lengths from 50 to 150 and find that models with $n = 100$ and $n = 125$ achieve a strong mIoU of 65.7%. We also examine different query dimensions and observe optimal performance at $c = 256$. Notably, increasing query count or dimensionality introduces redundancy and noise, weakening semantic understanding, and also increases overfitting risk, making the model overly sensitive to minor details.

| Method | Learnable Params. (M) | Inference Time (s) | GFLOPs |
|---|---|---|---|
| CAT-Seg | 70.27 | 0.37 | 2000.57 |
| DEDOS (Ours) | **23.57** | **0.12** | **1030.47** |

Table 7. Training and testing efficiency comparison. All results are measured with a single RTX 4090 GPU. The resolution of the input image is 640 × 640. The clip model is ViT-L/14.

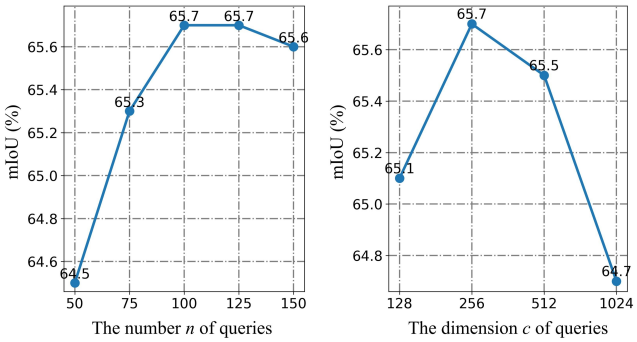


Figure 4. Ablation study on the number and the dimension of proxy queries.

**Efficiency comparison.** Table 7 provides a detailed efficiency comparison between our method and the SOTA approach CAT-Seg [11], evaluating learnable parameters, inference time, and GFLOPs. The results highlight our model's superior efficiency in both training and inference.

## 5. Conclusion

In this work, we propose DEDOS, a denoising learning framework based on the diffusion model for open-vocabulary semantic segmentation. DEDOS treats the image as a label embedded with "noise"—non-essential details for perceptual tasks—and employs proxy queries to progressively construct the scene skeleton during denoising process. Meanwhile, CLIP features guide the decoupling of diffusion features, preventing proxy queries from being disrupted by texture information. Proxy queries are ultimately used to enhance CLIP's scene understanding and serve as an interface for multi-level text-visual interactions. Extensive experiments demonstrate that DEDOS significantly outperforms previous SOTA methods across various benchmarks. Even in scenarios with vast differences from the training dataset, our method surpasses previous SOTA methods by a large margin, underscoring its robustness in real-world scenarios. This work opens up a new direction for effectively leveraging the internal representation of diffusion models to improve vision-language models.

# References

[1] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 2

[2] Yasser Benigmim, Subhankar Roy, Slim Essid, Vicky Kalogeiton, and Stéphane Lathuilière. Collaborating foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3108–3119, 2024. 2

[3] Benedikt Blumenstiel, Johannes Jakubik, Hilde Kühne, and Michael Vössing. What a mess: Multi-domain evaluation of zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 6, 7

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

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3

[6] 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. 1

[7] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023. 2

[8] Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, and David J Fleet. A generalist framework for panoptic segmentation of images and videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 909–919, 2023. 2

[9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 1

[10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 3, 5, 6, 8

[11] 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. 3, 6, 7, 8

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 2

[13] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11583–11592, 2022. 1, 6

[14] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 6

[15] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019. 1

[16] 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. 1, 2, 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[18] Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu, Zhenguo Li, and Ping Luo. Ddp: Diffusion model for dense visual prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21741–21752, 2023. 2

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1, 2

[20] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023. 2

[21] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1780–1790, 2021. 3

[22] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020. 5

[23] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 2

[24] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2

[25] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1, 2, 6

[26] Yuhe Liu, Chuanjian Liu, Kai Han, Quan Tang, and Zengchang Qin. Boosting semantic segmentation from the perspective of explicit class embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 821–831, 2023. 3

[27] 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. 1

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[29] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 2

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2

[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. 6

[32] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6, 8

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[35] 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. 1, 2, 7

[36] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[37] 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. 7

[38] 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. 6

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[40] Yuan Wang, Rui Sun, Naisong Luo, Yuwen Pan, and Tianzhu Zhang. Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3952–3963, 2024. 1

[41] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28619–28630, 2024. 3

[42] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. *Advances in Neural Information Processing Systems*, 36:54683–54695, 2023. 2

[43] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023. 2

[44] 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. 6

[45] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2, 6, 7

[46] 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. 6

[47] 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. 1, 2, 3, 6

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

[49] Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, and Hongtao Lu. A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7071–7080, 2023. 1, 2

[50] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 6

[51] Zilong Zhong, Zhong Qiu Lin, Rene Bidart, Xiaodan Hu, Ibrahim Ben Daya, Zhifeng Li, Wei-Shi Zheng, Jonathan Li, and Alexander Wong. Squeeze-and-attention networks for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13065–13074, 2020. 1

[52] 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. 6

[53] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1, 2, 3

[54] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 1, 2, 3, 6

[55] 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. 6