Diffusion Model as a Generalist Segmentation Learner

Haoxiao Wang*,†,1, Antao Xiang*,2, Haiyang Sun*,1, Peilin Sun3, Changhao Pan1, Yifu Chen1, Minjie Hong1, Weijie Wang1, Shuang Chen1, Yue Chen4, and Zhou Zhao‡,1

[1] Zhejiang University [2] South China University of Technology
[3] Nanjing University [4] Peking University
Project Page: https://wang-haoxiao.github.io/DiGSeg/
{haoxiao.wang, zhaozhou}@zju.edu.cn

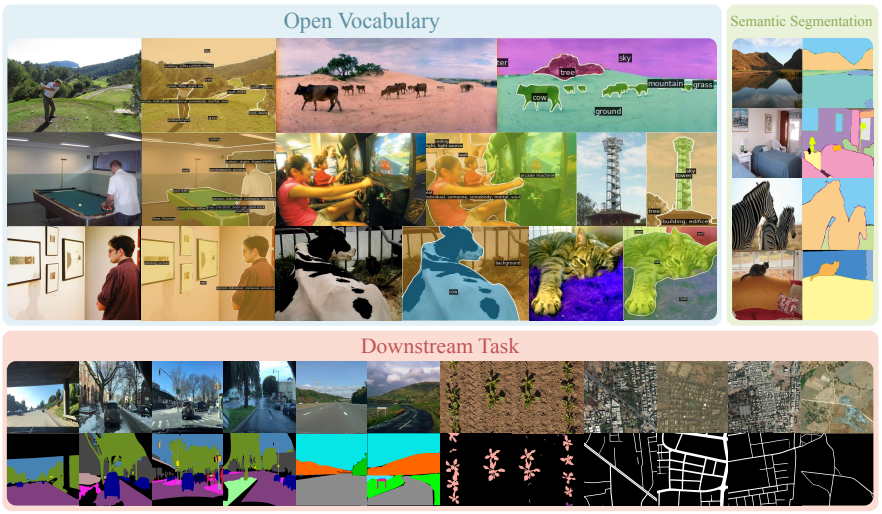

Fig. 1: We introduce DiGSeg, a general-purpose segmentation framework built on pretrained diffusion models. By exploiting the strong spatial priors encoded in generative models and fine-tuning with segmentation-aware objectives, DiGSeg produces consistent, high-quality masks across diverse settings, including semantic segmentation, open-vocabulary queries, and cross-domain datasets spanning medical, agricultural, and remote sensing imagery.

Abstract. Diffusion models are primarily trained for image synthesis, yet their denoising trajectories encode rich, spatially aligned visual priors. In this paper, we demonstrate that these priors can be utilized for text-conditioned semantic and open-vocabulary segmentation, and this approach can be generalized to various downstream tasks to make a general-purpose diffusion segmentation framework. Concretely, we introduce DiGSeg (**Di**ffusion Models as a **G**eneralist **Seg**mentation Learner), which repurposes a pretrained diffusion model into a unified segmentation framework. Our approach encodes the input image and ground-truth

* Equal contribution, † Project Leader, ‡ Corresponding author.

mask into the latent space and concatenates them as conditioning signals for the diffusion U-Net. A parallel CLIP-aligned text pathway injects language features across multiple scales, enabling the model to align textual queries with evolving visual representations. This design transforms an off-the-shelf diffusion backbone into a universal interface that produces structured segmentation masks conditioned on both appearance and arbitrary text prompts. Extensive experiments demonstrate state-of-the-art performance on standard semantic segmentation benchmarks, as well as **strong open-vocabulary generalization** and **cross-domain transfer** to medical, remote sensing, and agricultural scenarios—without domain-specific architectural customization. These results indicate that modern diffusion backbones, can serve as generalist segmentation learners rather than pure generators, narrowing the gap between visual generation and visual understanding.

1 Introduction

Modern segmentation systems have achieved remarkable progress across a wide range of tasks—from semantic and instance segmentation in natural scenes to highly specialized domains such as medical imaging, remote sensing, and agriculture [32,93,107]. Despite this progress, the ecosystem remains fundamentally fragmented. Most models are tailored to a specific task or domain, relying on different architectures, label spaces, and training pipelines [30,45,70,71,80,101]. As a result, transitioning from closed-vocabulary semantic segmentation to open-vocabulary recognition, or from natural images to aerial or medical imagery, often requires redesigning large portions of the system [46,82,95]. This fragmentation highlights a core question: can we design a single, unified segmentation model capable of operating robustly across tasks, vocabularies, and visual domains?

Recent research indicates that diffusion models may provide a pathway to a unified segmentation framework. Several approaches utilize pretrained text-to-image diffusion models by extracting their cross- or self-attention maps to create segmentation masks [1, 3, 10, 33, 42, 76, 78, 99, 108]. These studies highlight an intriguing characteristic: diffusion backbones encode rich semantic correspondences that naturally align visual regions with textual or structural cues. DiffSeg [73] aggregates self-attention maps; DiffCut [16] segments images by employing diffusion features in conjunction with graph-based clustering; and DiffuMask [85] leverages cross-attention to synthesize images and masks simultaneously. Collectively, these methods illustrate that diffusion models organize visual concepts across spatial dimensions, making them a promising foundation for more comprehensive segmentation learning systems.

Despite the potential of diffusion-based repurposing methods, they often fail to deliver reliable, high-quality segmentation results. Many of these methods depend on raw attention maps, which tend to be noisy, low-resolution, and inconsistent across different layers. As a result, they frequently produce fragmented masks that require significant post-processing to be usable [5, 73]. While high-

resolution attention maps can provide more detail, they often lack coherence. In contrast, low-resolution maps offer semantic consistency, but at the expense of losing information about small objects and fine boundaries. Further analysis of diffusion transformers reveals that although some layers exhibit strong semantic grounding, effectively utilizing them for dense prediction remains a challenge [39]. Moreover, existing methods generally focus on a single task—such as open-vocabulary segmentation, panoptic parsing, or instance segmentation—and do not demonstrate that a single diffusion backbone can effectively generalize across multiple tasks and visual domains [28, 40, 42, 90]. In other words, the field currently lacks a unified and conditioned interface that can convert a pretrained diffusion model into a comprehensive segmentation engine.

In this paper, we make that step. Our main idea is to transform this implicit capability into an explicit segmentation interface using a straightforward fine-tuning protocol. We encode an RGB image along with its ground-truth segmentation map into the latent space of a pretrained diffusion model. In this process, we keep almost all components frozen and fine-tune only the denoising U-Net to generate denoising outputs that align with the segmentation-consistent latents. This approach retains the powerful visual prior of the generator sourced from a vast internet-scale dataset. Unlike attention-based methods, which rely on post-processing to create masks, our method directly teaches the model to produce high-quality, text-controllable masks from supervised data. Consequently, we develop a diffusion model that functions effectively as a segmentation model rather than just a generator that sometimes produces usable masks.

Our main contributions are as follows:

- We present a novel and effective method for fine-tuning a pretrained diffusion model into a segmentation model, showcasing the potential of diffusion models for segmentation tasks.
- We introduce a visual latent pathway plus a CLIP-aligned text conditioner that injects language at multiple denoising scales, giving the diffusion U-Net an explicit mechanism to bind textual queries to evolving mask latents.
- Our extensive experiments show state-of-the-art performance on standard segmentation tasks, strong zero-shot and open-vocabulary capabilities, and excellent cross-domain transfer without task-specific architectures, proving the model as a generalist segmentation learner.

2 Related Work

2.1 Standard Segmentation Tasks

Image segmentation is a fundamental task in computer vision aimed at dividing an image into pixel-level segments based on semantics [7, 12, 30, 41, 69]. This involves providing a segmentation mask and a class label for each segment. In semantic segmentation, the goal is to output a single segment for each class present in the image [35, 65]. Traditional architectures such as DeepLabv3+ [8] demonstrate the strong ability of CNNs to capture multi-scale context and achieve high

accuracy in semantic segmentation benchmarks. Recently, transformer-based frameworks have further advanced the field. Mask2Former [12] adopts a masked-attention transformer decoder to address panoptic, instance, and semantic segmentation in a unified architecture. OneFormer [34] further extends this idea by introducing task-conditioned joint training, enabling a single model to handle all segmentation tasks simultaneously. Despite these advances, both CNN- and transformer-based models still suffer from limited generalization. They are restricted to finite, dataset-specific vocabularies—far smaller than the rich concepts used to describe the real world—and thus struggle to recognize or segment unseen categories without retraining.

2.2 Open-Vocabulary Segmentation

Open-vocabulary segmentation aims to recognize and segment arbitrary categories beyond the training set by leveraging language supervision [19, 26, 26, 27, 44,47,54,89,97]. Early methods [4,86,100]in this area focus on aligning visual features with pre-trained text embeddings by learning a feature mapping that associates visual and text spaces effectively. With the advent of vision–language models such as CLIP [60], open-vocabulary segmentation has emerged as a promising direction to overcome the limited label spaces of traditional models. CLIP-based methods like OpenSeg [26], ZegFormer [18], and ZSseg [92] typically adopt a two-stage design: class-agnostic masks are first proposed and then matched to CLIP text embeddings for classification. While ODISE [90] leverages diffusion models to produce high-quality masks, it still relies on region proposal networks trained with limited annotations and prompt-based category matching. In contrast, our model integrates segmentation awareness and text conditioning directly within the diffusion process, removing the need for proposals or handcrafted prompts and achieving stronger generalization across unseen categories and domains.

2.3 Diffusion Models for Segmentation

Recent research has explored the potential of using generative models, particularly diffusion models [31], for segmentation tasks [1,3,9,42,76,78,99,108]. Early efforts, such as DiffuseSeg [73], DiffCut [80], DiffuMask [85], and Seg4Diff [39], repurposed pretrained text-to-image diffusion models by extracting their internal attention maps or latent features. These approaches demonstrated that diffusion backbones inherently encode strong spatial and semantic organization. However, they generally rely on attention maps from intermediate layers, which are often low-resolution, noisy, and inconsistent, resulting in fragmented masks. This leads to a heavy dependency on heuristic post-processing techniques. In contrast, we propose a more direct and unified approach. Rather than interpreting diffusion attention post-hoc, we fine-tune the denoising U-Net to explicitly produce segmentation-consistent latents. This design effectively teaches diffusion to segment, generating high-quality, text-controllable masks without the need for manual post-processing. Our framework recasts the diffusion model from serving

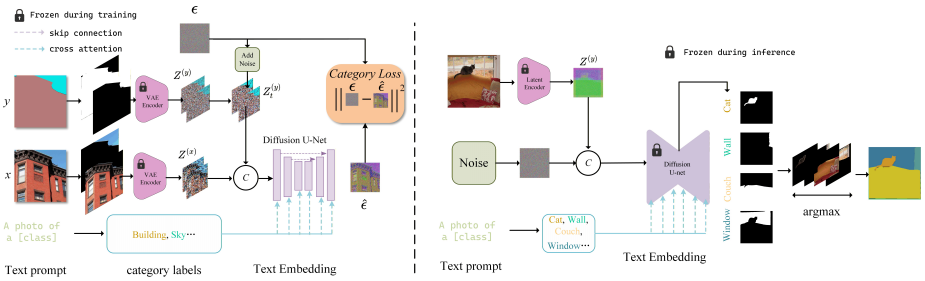

Fig. 2: DiGSeg pipeline overview, which presents the training and inference pipelines of our diffusion-based generation model. In training, paired images are encoded into latent space and, together with text prompts, guide the diffusion U-Net to predict noise using an MSE objective. In inference, noise is sampled and progressively denoised under text conditioning, after which the VAE decoder reconstructs the final image.

solely as a generative prior to functioning as a generalized segmentation learner, preserving its rich visual knowledge while enabling explicit semantic control.

2.4 Diffusion Models for Dense Prediction

Akin to segmentation, diffusion models have recently shown great promise in other dense prediction tasks, most notably in monocular depth estimation [11, 20, 22, 37, 57, 63, 64, 74, 77]. DiffusionDepth [20] firstly transforms the monocular depth estimation task into an iterative denoising diffusion process guided by image features. Marigold [37] repurposes pretrained image diffusion models to predict continuous depth maps by fine-tuning the denoising U-Net on large-scale depth datasets. While both DiGSeg and these depth-oriented models leverage the strong spatial priors of generative backbones, they differ in several key aspects. First, depth estimation typically focuses on predicting a single-channel continuous value from visual cues alone, whereas DiGSeg is designed as a general-purpose model that handles diverse discrete label spaces—ranging from closed-vocabulary semantic classes to arbitrary open-vocabulary text queries. Second, unlike those depth estimation models which are primarily image-conditioned, DiGSeg introduces a CLIP-aligned text pathway that enables explicit language-visual grounding at multiple denoising scales. This allows our model to not only capture geometric structures but also to align semantic concepts across diverse domains such as medical and remote sensing imagery, a capability not addressed in specialized depth diffusion models.

3 Method

3.1 Problem Definition

Segmentation tasks can be categorized into three main types: semantic segmentation, open-vocabulary segmentation, and domain-specific segmentation. Seman-

tic segmentation involves a fixed set of labels, denoted as $\mathcal{C}_{\text{train}}$, where the goal is to predict a class ID for each pixel in an image. Open-vocabulary segmentation expands upon $\mathcal{C}_{\text{train}}$ to include unseen categories, represented as $\mathcal{C}_{\text{test}} \neq \mathcal{C}_{\text{train}}$, by utilizing image-text alignment techniques [26, 75, 90]. Domain-specific segmentation, which can include applications such as medical imaging or remote sensing, often depends on specialized architectures. This reliance can restrict the ability to generalize across different domains.

3.2 Method Overview

Figure 2 presents an overview of our DiGSeg framework. Inspired by previous works [37, 90], We found that fine-tuning the diffusion model yields better results than treating the diffusion model as a feature extractor for the segmentation task. We repurpose a pretrained diffusion model into a unified segmentation learner capable of semantic and open-vocabulary segmentation across diverse domains. The system contains three key components. First, the **Visual Latent Pathway** (Sec. 3.4) encodes the RGB image and its segmentation map into a compact latent space using the Stable Diffusion VAE. This preserves spatial structure while enabling efficient learning of fine-grained correspondences. Second, the **CLIP-Aligned Text Conditioning Module** (Sec. 3.5) injects language features across denoising steps via a frozen CLIP text encoder, enabling open-vocabulary reasoning without task-specific prompts or additional heads. Finally, the **Segmentation-Consistent Denoising U-Net** (Sec. 3.6) is trained to denoise toward segmentation-consistent latents conditioned on both image and text features, allowing the diffusion backbone to directly generate segmentation maps instead of relying on attention-based heuristics. Once trained, DiGSeg supports open-vocabulary inference by conditioning on arbitrary text inputs and generalizes robustly to unseen categories and domains. The following sections detail each component.

3.3 Generative Formulation

We formulate segmentation as a conditional denoising diffusion generation task. Given an RGB image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times 3}$ and its corresponding segmentation map $\boldsymbol{y} \in \mathbb{R}^{H\times W}$, our goal is to model the conditional distribution $p_\theta(\boldsymbol{y}|\boldsymbol{x})$ within the latent space of a pretrained diffusion model.

Following the standard diffusion framework, we define a *forward* noising process that gradually corrupts the ground-truth segmentation latent $\boldsymbol{y}_0 := \boldsymbol{y}$ with Gaussian noise:

$$y_t = \sqrt{\bar{\alpha}_t}\, y_0 + \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ and $\{\beta_t\}_{t=1}^{T}$ denotes the variance schedule over T diffusion steps. The reverse process is parameterized by a denoising network $\boldsymbol{\epsilon}_\theta(\cdot)$, a U-Net, which learns to progressively remove noise from y_t conditioned on x:

$$p_\theta(y_{t-1}|y_t, x) = \mathcal{N}\big(y_{t-1}; \mu_\theta(y_t, x, t), \Sigma_\theta(y_t, x, t)\big). \tag{2}$$

During training, parameters θ are optimized using the standard diffusion objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{y_0, \boldsymbol{\epsilon}, t}\big[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(y_t, x, t)\|_2^2\big], \tag{3}$$

Where $\hat{\boldsymbol{\epsilon}} = \boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)$ is the noise estimate. Unlike prior works that model $p_\theta(x|y)$ for image generation, our formulation reverses the conditioning to directly generate segmentation-consistent latents y from the image $\boldsymbol{x}$. At inference time, the segmentation latent $\boldsymbol{y} := \boldsymbol{y}_0$ is reconstructed starting from a normally distributed variable $\boldsymbol{y}_T$, by iteratively applying the learned denoiser $\boldsymbol{\epsilon}_\theta(\boldsymbol{y}_t, \boldsymbol{x}, t)$. Conditioned on the input image $\boldsymbol{x}$, the model gradually refines $\boldsymbol{y}_t$ through the reverse diffusion process to produce a segmentation-consistent latent representation, which can be decoded into the final mask prediction $\boldsymbol{y}_0$.

3.4 Visual Latent Pathway

The encoder $\mathcal{E}$ compresses both the RGB image and its corresponding segmentation map into compact latent representations:

$$z_x = \mathcal{E}(x), \quad z_y = \mathcal{E}(y),$$

where $z_x, z_y \in \mathbb{R}^{h \times w \times c}$ denote the encoded latent features of the image and the segmentation map, respectively. The decoder $\mathcal{D}$ allows reconstruction back to the pixel space, i.e., $\hat{x} = \mathcal{D}(z_x)$ and $\hat{y} = \mathcal{D}(z_y)$, ensuring perceptual consistency between latent and data domains.Because the pretrained VAE is designed for 3-channel RGB inputs, the single-channel segmentation map is replicated across three channels to simulate an RGB image before encoding. This simple strategy maintains compatibility with the encoder without retraining or modifying the latent structure.

3.5 Text Conditioner

To endow our model with open-vocabulary and text-controllable segmentation capability, we introduce a CLIP-Aligned Text Conditioner that injects language information into the diffusion process. At each denoising timestep, this module provides semantic grounding by aligning textual and visual representations within the U-Net.

Given a class name or a natural-language description, we obtain a text embedding t_{clip} using a frozen CLIP text encoder. This embedding is then integrated into multiple denoising scales of the U-Net through cross-attention:

$$h_t = \text{UNet}(z_t, t_{\text{clip}}, t), \tag{4}$$

where z_t denotes the noisy latent at timestep t. Multi-scale language injection ensures that both global semantics and local spatial cues are jointly refined during denoising.

3.6 Segmentation-Consistent Denoising

Given the encoded latent pairs (z_x, z_y) from the Visual Latent Pathway, we randomly sample a timestep $t \in [1, T]$ and add Gaussian noise to the segmentation latent z_y according to the forward diffusion process:

$$z_y^t = \sqrt{\bar{\alpha}_t}\, z_y + \sqrt{1 - \bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{5}$$

where $\bar{\alpha}_t$ follows the predefined noise schedule. The denoiser, implemented as a U-Net ϵ_θ, then learns to predict and remove this noise conditioned on both the image latent z_x and the text embedding t_{clip}:

$$\mathcal{L}_{\text{seg}} = \mathbb{E}_{t,\epsilon}\big[\|\epsilon - \epsilon_\theta(z_y^t, z_x, t, t_{\text{clip}})\|_2^2\big]. \tag{6}$$

Noise Strategy. To stabilize training and accelerate convergence, we employ an *annealed multi-scale noise schedule* [68, 84]. Instead of adding noise purely at a single resolution, we combine Gaussian noise components at multiple spatial scales, each upsampled to match the latent resolution. At early diffusion steps, high-frequency perturbations dominate to encourage fine-structure learning; as denoising progresses, low-frequency components become more influential, gradually emphasizing semantic structure. This annealed multi-scale perturbation improves spatial coherence and yields smoother, more accurate segmentation boundaries.

3.7 Inference

At inference time, our model performs conditional latent diffusion denoising to generate segmentation maps from an input image. Given an RGB image x, we first encode it into the latent space as $z_x = \mathcal{E}(x)$. A segmentation latent z_y^T is then initialized as standard Gaussian noise $z_y^T \sim \mathcal{N}(0, I)$ and progressively denoised under the same schedule used during training:

$$z_y^{t-1} = \frac{1}{\sqrt{\alpha_t}}\Big(z_y^t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\, \epsilon_\theta(z_y^t, z_x, t, t_{\text{clip}})\Big) + \sigma_t \epsilon, \tag{7}$$

where σ_t controls the noise level and ϵ_θ denotes our segmentation-consistent denoiser. After completing all reverse steps, the clean segmentation latent z_y^0 is decoded back to the pixel space using the frozen VAE decoder:

$$\hat{y} = \mathcal{D}(z_y^0). \tag{8}$$

Test-Time Ensembling. Given the stochastic nature of the diffusion process, we optionally adopt a lightweight test-time ensemble: multiple inference passes are run with different noise seeds, and the resulting segmentation maps are averaged in the latent space before decoding. This aggregation improves spatial consistency and reduces noise artifacts, particularly in fine-structure regions, with minimal additional computational cost.

Method	VLM	Backbone	Training Dataset	Additional Dataset	A-847	PC-459	A-150	PC-59	Cityscapes
ODISE [90]	CLIP ViT-L/14	Stable Diffusion	COCO-Panoptic	✗	11.0	13.8	28.7	55.3	-
OVSeg [49]	CLIP ViT-L/14	Swin-B	COCO-Stuff	✓	9.0	12.4	29.6	55.7	-
SAN [91]	CLIP ViT-L/14	Side Adapter	COCO-Stuff	✗	12.4	15.7	29.9	51.8	-
SCAN [51]	CLIP ViT-L/14	Swin-B	COCO-Stuff	✗	14.0	16.7	33.5	59.3	-
CAT-Seg [14]	CLIP ViT-L/14	-	COCO-Stuff	✗	16.0	23.8	31.5	62.0	-
MAFT+ [36]	ConvNeXt-L	-	COCO-Stuff	✗	15.1	21.6	36.1	59.4	-
SED [87]	CLIP ConvNeXt-L	-	COCO-Stuff	✗	13.9	22.6	35.2	60.6	-
Mask-Adapter [48]	CLIP ConvNeXt-L	-	COCO-Stuff	✗	16.2	22.7	38.2	60.4	37.9
Seg4Diff [39]	CLIP ViT-L/14	Stable Diffusion	COCO-Stuff	✗	-	-	35.2	51.2	26.0
HyperCLIP [58]	CLIP ViT-L/14	-	COCO-Stuff	✗	16.3	24.1	38.2	64.2	-
OVSNet [52]	CLIP ViT-L/14	ResNet-101c	COCO-Stuff	✗	16.2	23.5	37.1	62.0	-
DPSeg [102]	CLIP ConvNeXt-L	-	COCO-Stuff	✗	15.7	24.1	37.1	62.3	-
SemLA [59]	CLIP ConvNeXt-L	-	COCO-Stuff	✗	-	-	36.9	62.2	-
ESC-Net [43]	CLIP ViT-L/14	-	COCO-Stuff	✗	18.1	27.0	41.8	65.6	-
DiGSeg (ours)	CLIP ViT-L/14	Stable Diffusion	COCO-Stuff	✗	**19.9** (+1.8)	**29.2** (+2.2)	**43.2** (+1.4)	**68.4** (+2.8)	**38.5** (+0.6)
OVSeg [49]	CLIP ViT-B/16	ResNet-101c	COCO-Stuff	✓	7.1	11.0	24.8	53.3	-
SCAN [51]	CLIP ViT-B/16	Swin-B	COCO-Stuff	✗	10.8	13.2	30.8	58.4	-
EBSeg [65]	CLIP ViT-B/16	SAM ViT-B	COCO-Stuff	✗	11.1	17.3	30.0	56.7	-
SED [87]	ConvNeXt-B	-	COCO-Stuff	✗	11.4	18.6	31.6	57.3	-
CAT-Seg [14]	CLIP ViT-B/16	-	COCO-Stuff	✗	12.0	19.0	31.8	57.5	-
OPMapper [81]	CLIP ViT-B/16	Swin-B	COCO-Stuff	✗	-	-	31.0	58.3	-
ESC-Net [43]	CLIP ViT-B/16	-	COCO-Stuff	✗	13.3	21.1	35.6	59.0	-
HyperCLIP [58]	CLIP ViT-B/16	-	COCO-Stuff	✗	12.3	19.2	32.1	58.5	-
Mask-Adapter [48]	CLIP ConvNeXt-B	-	COCO-Stuff	✗	14.2	17.9	35.6	58.4	35.2
DPSeg [102]	CLIP ConvNeXt-B	-	COCO-Stuff	✗	12.5	20.1	33.3	58.4	-
DiGSeg (ours)	CLIP ViT-B/16	Stable Diffusion	COCO-Stuff	✗	**17.5** (+3.3)	**23.1** (+2.0)	**37.2** (+1.6)	**62.7** (+3.7)	**36.5** (+1.3)

Table 1: Quantitative evaluation on open-vocabulary segmentation benchmarks All methods are evaluated on A-847, PC-459, A-150, PC-59, and Cityscapes. We highlight the **best**, while the second-best results are underlined. Improvements over the second-best are highlighted in green.

Hyperparameter τ Tuning. Since the output of our diffusion model is a continuous-valued mask in $[0, 1]$, a threshold τ is required to convert the predicted logits into a binary mask for each class. The choice of τ can influence the sharpness and coverage of the resulting mask, and different categories may exhibit different optimal threshold values depending on object size, texture, and spatial sparsity. To characterize this behavior, we examine how IoU varies with respect to τ for several representative categories on Fig. 3. The optimal threshold can shift across classes, with small or fine-grained objects typically favoring slightly lower thresholds, while larger and more homogeneous objects prefer higher thresholds. We intentionally avoid post-processing to preserve the generality and simplicity of our model. Our empirical findings indicate that a single fixed value of $\tau = 0.7$ consistently delivers strong performance across semantic segmentation, open-vocabulary segmentation, and all downstream segmentation benchmarks.

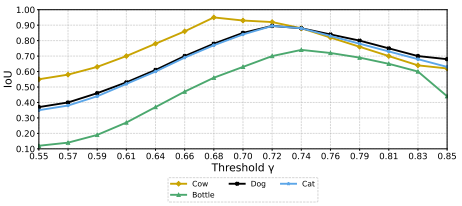

Fig. 3: Impact of Hyperparameter τ.

Method	COCO		ADE20K	
	input size	miou	input size	miou
SegFormer-B5 [88]	512^2	46.7	640^2	51.8
Mask2Former-Swin-L [12]	-	-	640^2	<u>57.3</u>
OneFormer [34]	-	-	640^2	57.0
LDMSeg [76]	-	-	512^2	52.2
PEM [6]	-	-	512^2	45.5
FeedFormer-B2 [66]	-	-	512^2	48.0
CGRSeg-L [56]	512^2	46.0	512^2	48.3
VWFormer-B5 [94]	512^2	48.0	512^2	54.7
SegMAN [23]	512^2	48.2	512^2	53.2
EoMT [38]	512^2	<u>48.7</u>	512^2	57.1
OffSeg-L [98]	512^2	46.0	512^2	48.5
MambaVision-B [29]	-	-	512^2	49.1
DiGSeg (ours)	512^2	**50.8** (+2.1)	512^2	**58.6** (+1.3)

Table 2: Quantitative evaluation on semantic segmentation. Our DiGSeg demonstrates outstanding performance.

Method	IoU_{road}	Precision	Recall	F1
DDCTNet [24]	64.27	79.02	78.10	78.24
Unet [62]	62.94	-	-	-
D-LinkNet [106]	63.00	-	-	-
CoANet [53]	60.65	76.98	72.34	74.61
SGCN [105]	53.92	72.95	68.25	72.31
DeepLabv3 [82]	61.97	77.26	74.09	75.64
Segroad [72]	66.23	-	-	-
CGC-Net [103]	68.80	82.67	80.39	81.51
SegMAN	48.12	70.25	68.10	69.16
EoMT	52.52	72.88	71.35	72.10
OffSeg	51.75	72.10	70.85	71.46
MambaVision	<u>57.28</u>	<u>75.32</u>	<u>74.50</u>	<u>74.91</u>
DiGSeg (ours)	**65.78** (+8.50)	**79.93** (+4.61)	**78.92** (+4.42)	**78.79** (+3.88)

Table 3: Quantitative evaluation on DeepGlobe road segmentation tasks. The module specifically designed for remote sensing tasks is denoted by gray.

4 Experiments

We begin this section by describing the implementation details of our framework. We then compare our method with SOTA approaches on semantic and open-vocabulary segmentation benchmarks. Finally, we conduct comprehensive ablation studies to validate the contribution of each component in our model. Due to page constraints, more implementation details, dataset tests, detailed experimental results and metrics, as well as ablation experiment analyses, are presented in the **supplementary materials**.

4.1 Implementation Details

Architecture We implement DiGSeg utilize SDv2 [61]. We adopt the DDPM noise schedule [31] with $T = 1000$ steps during training. For each training iteration, we sample a timestep $t \sim \mathcal{U}(\{1, \ldots, T\})$. At inference time, we use a DDIM sampler [67] with between 1 and 50 steps. For all segmentation tasks, unless otherwise stated, we ensemble 8 predictions from different initial noise. we resize the input images to 512×512, use random horizontal flip and scale jittering in $[0.8, 1.2]$. The model is trained for 30000 iterations with a total batch size of 16. We use AdamW with a learning rate of 1×10^{-4}. Training and testing are conducted using 8 NVIDIA A100 GPU, For most semantic segmentation and open vocabulary segmentation tasks, it takes approximately 1 days to converge during training.

Training datasets We use widely adopted benchmarks for image segmentation. Specifically, we evaluate the semantic segmentation performance on COCO [50]

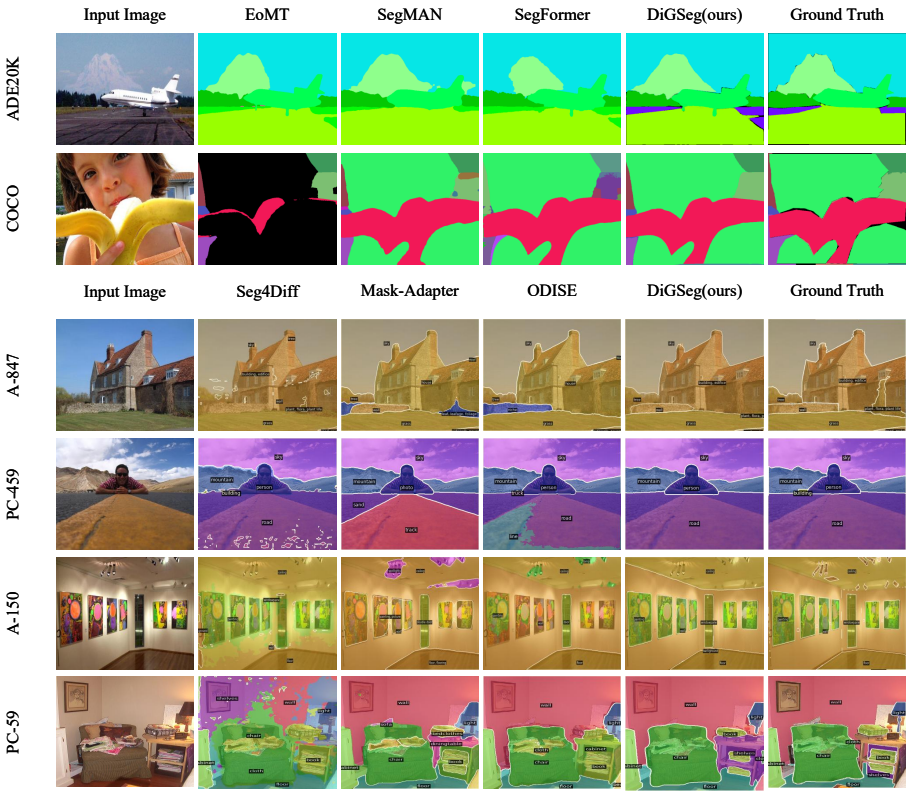

Fig. 4: Qualitative comparison of semantic and open-vocabulary segmentation across different datasets.

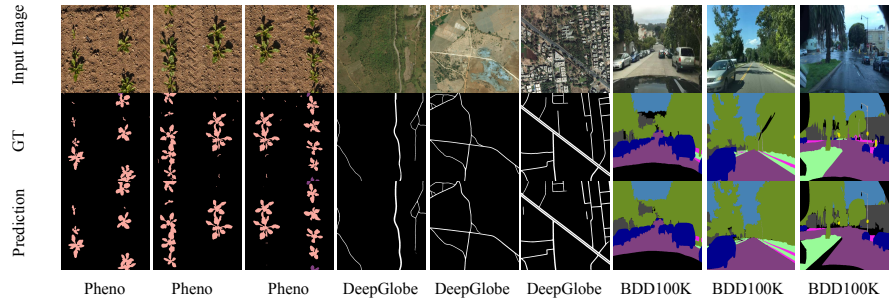

Fig. 5: Qualitative results of cross-domain segmentation across different datasets.

and ADE20K [104]. Following the experiment in [26, 91], we report performance on A-150 with 150 common classes and A-847 with all the 847 classes of ADE20K [104], PC-59 with 59 common classes and PC-459 with full 459 classes of Pascal Context [55], we also incorporated the CityScapes [15] dataset in order

to conduct a more comprehensive comparison. For task-specific segmentation tasks, we conducted extensive experiments using various datasets from medical, remote sensing, and agricultural domains. These datasets include Agriculture-Vision [13], REFUGE-2 [21], BraTS-2021 [2], DeepGlobe [17], LoveDA [79], and BDD100K [96]. Due to page constraints, we have focused on presenting the Phenobench [83] and DeepGlobe [17] datasets in the main sections.

Evaluation metrics For semantic and open-vocabulary segmentation, we evaluate performance using the mean Intersection-over-Union (mIoU) across all classes, following the official dataset splits. For all task-specific segmentation benchmarks, mIoU is used as the primary metric. In the medical segmentation setting, we additionally report the Dice score to better capture region overlap quality.

4.2 Comparison with State of the Art

Semantic and open-vocabulary segmentation We compare DiGSeg with state-of-the-art segmentation frameworks in several segmentation tasks. As shown in Tab. 1, for open-vocabulary segmentation, we compare DiGSeg to thirteen baselines on mIoU score. For almost the datasets, we achieved state-of-the-art results. Notebly, We adopt the public results of integrating Mask-Adapter [48] into MAFTP [36] as the baseline (i.e., 'MAFTP w/ MaskAdapter'), similarly, SemLA [59] is based on CAT-Seg [14], and OPMapper [81] is based on SCAN [51]. The results of semantic segmentation are presented in Tab. 2. For all baseline methods, we select the variant with the largest number of parameters to ensure a fair comparison. Our proposed DiGSeg consistently outperforms existing segmentation frameworks on both the COCO and ADE20K benchmarks. Specifically, DiGSeg achieves an mIoU of **50.8** on COCO and **58.6** on ADE20K, exceeding the second-best results by $+$**2.1** and $+$**1.3**, respectively.

Task-specific segmentation The quantitative results for the DeepGlobe dataset can be found in Table 3, while the results for the Monuseg dataset are shown in supplementary materials. To demonstrate the generalized capabilities of our model, all settings are consistent with those used for semantic segmentation. Notably, our model demonstrates a significant performance in the general domains as well as in agriculture and remote sensing whitout any change in architecture. Additionally, the scores in medical domains are relatively low, likely reflecting CLIP's limited understanding of this specialized field.

4.3 Ablation Study

Data efficiency under limited supervision. To examine data efficiency and robustness to limited supervision, we train our model with 1/2, 1/4, 1/8 and 1/16 of the ADE20K training data. Figure 6 shows that our model can achieve nearly identical performance using only half the amount of data compared to the full dataset. Furthermore, when trained with just a quarter of the data, the results are still impressively strong. This shows that our diffusion-based

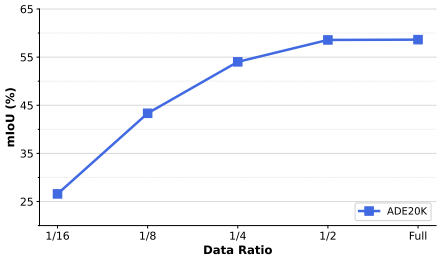

Fig. 6: Effect of training data ratio. Our model exhibits remarkable data efficiency, maintaining nearly equivalent performance even when trained on only 50% of the total dataset.

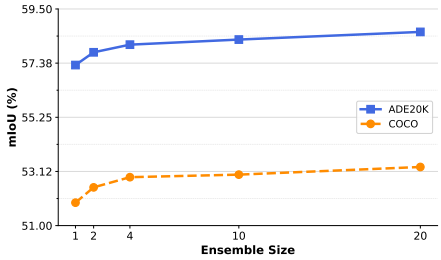

Fig. 7: Ablation of ensemble size. We observe a consistent improvement as the ensemble size increases. However, this improvement begins to taper off after 10 predictions per sample.

segmentation learner effectively utilizes pretrained visual priors and acquires generalizable representations in low-data scenarios.

Test-time ensembling. We investigated the impact of the number of test-time ensembling iterations on the outcome. As shown in Figure 7, a single prediction has already been able to yield reasonably good results. Integration ten times will result in a 1.2% improvement on ADE20K dataset and 1.9% improvement on COCO dataset.

Denoising steps. In dense prediction scenarios, [25] observed that the conventional denoising diffusion implicit models (DDIM) leading time-step strategy causes inconsistencies between the training and inference processes. Based on this finding, we will switch to using trailing timesteps for the DDIM. This change has significantly improved our model's efficiency, with performance reaching saturation after just 1 DDIM step. Figure 8 shows the effect of applying DDIM-trailing on the open-vocabulary benchmark.

Training noise. We compared the effects of four different types of noise on the training process. As shown in Table 4, we discovered that incorporating multi-resolution noise alone led to a significantly greater improvement compared to using standard Gaussian noise. In contrast, adding annealing noise by itself had a relatively minor effect when applied to standard Gaussian noise. The best results were achieved by applying annealing noise after introducing multi-resolution noise.

Cross-domain analysis. We compare models trained on COCO, ADE20K, and both datasets by evaluating them on Cityscapes and BDD100K. As shown in Table 5, ADE20K training yields the highest mIoU on both targets. We conduct a systematic analysis of the number of ensembles (E) and DDIM denoising steps (S), denoted jointly as the E-Step setting ($E \times S$). This suggests that ADE20K offers greater scene diversity and more detailed pixel-level annotations, which effectively capture the spatial context and semantics of urban environments. These results confirm that more data or mixed datasets do not necessarily improve cross-domain performance—dataset relevance is crucial.

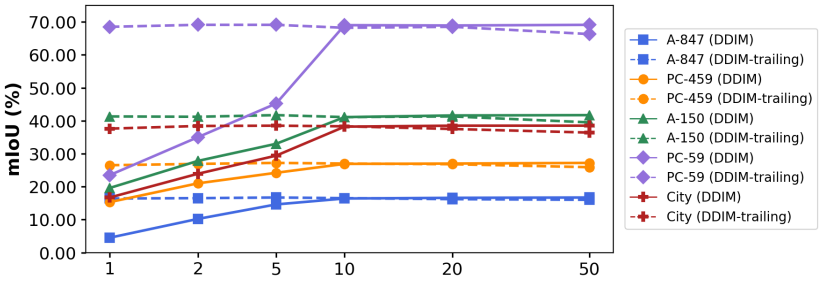

Fig. 8: Ablation study of denoising schedule. Under the DDIM-trailing setting, one step of denoising can achieve a relatively high effect, while DDIM requires at least 10 steps.

Method	COCO	ADE20K
Standard Gaussian noise	48.9	56.7
w/ Annealed noise	49.2	57.1
w/ Multi-resolution noise	49.7	57.6
w/ Multi-res. + Ann. (ours)	**50.8**	**58.6**

Table 4: Ablation on noise schedule. Combining multi-resolution and annealed noise yields best performance.

COCO	ADE20K	E-Step	Citys.	BDD100K
✗	✓	1×1	41.22	37.55
		8×2	**43.58**	**37.64**
✓	✗	1×1	37.80	35.76
		8×2	38.51	36.42
✓	✓	1×1	38.74	36.89
		8×2	39.04	37.21

Table 5: Cross-domain evaluation. Results on Cityscapes and BDD100K under different training sets.

Inference Speed. In Table 6, we present the inference runtime of DiGSeg. As a diffusion-based method, we admit our model is slower than feed-forward segmenters, which reflects the typical trade-off between quality and efficiency. However, many established diffusion acceleration techniques can be applied to further enhance the speed of future versions of our framework. In this work, we implement DDIM-trailing, which significantly improves upon standard diffusion sampling, making DiGSeg noticeably faster than previous diffusion-based segmentation models.

E-Step	COCO	ADE20K	FPS
1x1	48.2	56.8	11.27
1x2	48.5	57.1	10.61
4x2	50.5	58.5	5.82
8x2	50.8	58.6	3.15
20x50 (w/o trailing)	50.9	58.8	0.12

Table 6: Ablation on E-Step scheduling. Increasing the E-Step range improves segmentation accuracy but reduces inference speed.

5 Conclusion

In this paper, we presented **DiGSeg**, a novel framework that successfully repurposes pretrained diffusion models from pure generators into generalized segmentation learners. By introducing a latent-conditioned fine-tuning protocol and a multi-scale CLIP-aligned text pathway. Our extensive evaluations across

standard benchmarks and specialized domains demonstrate that DiGSeg not only achieves state-of-the-art performance but also exhibits exceptional open-vocabulary generalization. These results confirm that the rich visual priors within diffusion backbones can be effectively channeled through a unified interface, bridging the critical gap between generative modeling and dense visual understanding.

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
3. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
4. Bucher, M., Vu, T.H., Cord, M., Pérez, P.: Zero-shot semantic segmentation. Advances in Neural Information Processing Systems **32** (2019)
5. Cai, L., Zhao, K., Yuan, H., Zhang, Y., Zhang, S., Huang, K.: Freemask: Rethinking the importance of attention masks for zero-shot video editing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1898–1906 (2025)
6. Cavagnero, N., Rosi, G., Cuttano, C., Pistilli, F., Ciccone, M., Averta, G., Cermelli, F.: Pem: Prototype-based efficient maskformer for image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15804–15813 (2024)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
9. Chen, S., Shou, Q., Chen, H., Zhou, Y., Feng, K., Hu, W., Zhang, Y.F., Lin, Y., Huang, W., Song, M., et al.: Unify-agent: A unified multimodal agent for world-grounded image synthesis. arXiv preprint arXiv:2603.29620 (2026)
10. Chen, Y., Jiang, M., Zheng, K., Liang, J., Tie, C., Lu, H., Wu, R., Dong, H.: Learning part-aware dense 3d feature field for generalizable articulated object manipulation. arXiv preprint arXiv:2602.14193 (2026)
11. Chen, Y., Tie, C., Wu, R., Dong, H.: Eqvafford: Se (3) equivariance for point-level affordance learning. arXiv preprint arXiv:2408.01953 (2024)
12. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
13. Chiu, M.T., Xu, X., Wei, Y., Huang, Z., Schwing, A.G., Brunner, R., Khachatrian, H., Karapetyan, H., Dozier, I., Rose, G., et al.: Agriculture-vision: A large aerial image database for agricultural pattern analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2828–2838 (2020)

14. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4113–4123 (2024)
15. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
16. Couairon, P., Shukor, M., Haugeard, J.E., Cord, M., Thome, N.: Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. Advances in Neural Information Processing Systems **37**, 13548–13578 (2024)
17. Demir, I., Koperski, K., Lindenbaum, D., Pang, G., Huang, J., Basu, S., Hughes, F., Tuia, D., Raskar, R.: Deepglobe 2018: A challenge to parse the earth through satellite images. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 172–181 (2018)
18. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11583–11592 (2022)
19. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14084–14093 (2022)
20. Duan, Y., Guo, X., Zhu, Z.: Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In: European Conference on Computer Vision. pp. 432–449. Springer (2024)
21. Fang, H., Li, F., Wu, J., Fu, H., Sun, X., Son, J., Yu, S., Zhang, M., Yuan, C., Bian, C., et al.: Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. arXiv preprint arXiv:2202.08994 (2022)
22. Feng, Z., Kang, Z., Wang, Q., Du, Z., Yan, J., Shi, S., Yuan, C., Liang, H., Deng, Y., Li, Q., et al.: Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. arXiv preprint arXiv:2510.19400 (2025)
23. Fu, Y., Lou, M., Yu, Y.: Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19077–19087 (2025)
24. Gao, L., Zhou, Y., Tian, J., Cai, W.: Ddctnet: A deformable and dynamic cross-transformer network for road extraction from high-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–19 (2024)
25. Garcia, G.M., Abou Zeid, K., Schmidt, C., De Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 753–762. IEEE (2025)
26. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
27. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)
28. Gu, Z., Chen, H., Xu, Z.: Diffusioninst: Diffusion model for instance segmentation. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2730–2734. IEEE (2024)

29. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25261–25270 (2025)
30. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
31. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
32. Huang, L., Jiang, B., Lv, S., Liu, Y., Fu, Y.: Deep-learning-based semantic segmentation of remote sensing images: A survey. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 8370–8396 (2023)
33. Huang, S., Wu, J., Zhou, Q., Miao, S., Long, M.: Vid2world: Crafting video diffusion models to interactive world models. arXiv preprint arXiv:2505.14357 (2025)
34. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2989–2998 (2023)
35. Jain, J., Singh, A., Orlov, N., Huang, Z., Li, J., Walton, S., Shi, H.: Semask: Semantically masked transformers for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 752–761 (2023)
36. Jiao, S., Zhu, H., Huang, J., Zhao, Y., Wei, Y., Shi, H.: Collaborative vision-text representation optimizing for open-vocabulary segmentation. In: European Conference on Computer Vision. pp. 399–416. Springer (2024)
37. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
38. Kerssies, T., Cavagnero, N., Hermans, A., Norouzi, N., Averta, G., Leibe, B., Dubbelman, G., de Geus, D.: Your vit is secretly an image segmentation model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25303–25313 (2025)
39. Kim, C., Shin, H., Hong, E., Yoon, H., Arnab, A., Seo, P.H., Hong, S., Kim, S.: Seg4diff: Unveiling open-vocabulary segmentation in text-to-image diffusion transformers. arXiv preprint arXiv:2509.18096 (2025)
40. Kim, C., Ju, D., Han, W., Yang, M.H., Hwang, S.J.: Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15033–15042 (2025)
41. Kirillov, A., Girshick, R., He, K., Dollár, P.: Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6399–6408 (2019)
42. Le, M.Q., Nguyen, T.V., Le, T.N., Do, T.T., Do, M.N., Tran, M.T.: Maskdiff: Modeling mask distribution with diffusion probabilistic model for few-shot instance segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2874–2881 (2024)
43. Lee, M., Cho, S., Lee, J., Yang, S., Choi, H., Kim, I.J., Lee, S.: Effective sam combination for open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26081–26090 (2025)
44. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)
45. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmen-

tation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3041–3050 (2023)
46. Li, J., Cai, Y., Li, Q., Kou, M., Zhang, T.: A review of remote sensing image segmentation by deep learning methods. International Journal of Digital Earth **17**(1), 2328827 (2024)
47. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10965–10975 (2022)
48. Li, Y., Cheng, T., Feng, B., Liu, W., Wang, X.: Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14998–15008 (2025)
49. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7061–7070 (2023)
50. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
51. Liu, Y., Bai, S., Li, G., Wang, Y., Tang, Y.: Open-vocabulary segmentation with semantic-assisted calibration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3491–3500 (2024)
52. Liu, Y., Wu, S.L., Bai, S., Wang, J., Wang, Y., Tang, Y.: Stepping out of similar semantic space for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22664–22674 (2025)
53. Mei, J., Li, R.J., Gao, W., Cheng, M.M.: Coanet: Connectivity attention network for road extraction from satellite imagery. IEEE Transactions on Image Processing **30**, 8540–8552 (2021)
54. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European conference on computer vision. pp. 728–755. Springer (2022)
55. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 891–898 (2014)
56. Ni, Z., Chen, X., Zhai, Y., Tang, Y., Wang, Y.: Context-guided spatial feature reconstruction for efficient semantic segmentation. In: European Conference on Computer Vision. pp. 239–255. Springer (2024)
57. Patni, S., Agarwal, A., Arora, C.: Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28285–28295 (2024)
58. Peng, Z., Xu, Z., Zeng, Z., Wen, C., Huang, Y., Yang, M., Tang, F., Shen, W.: Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4562–4572 (2025)
59. Qorbani, R., Villani, G., Panagiotakopoulos, T., Colomer, M.B., Härenstam-Nielsen, L., Segu, M., Dovesi, P.L., Karlgren, J., Cremers, D., Tombari, F., et al.:

Semantic library adaptation: Lora retrieval and fusion for open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9804–9815 (2025)
60. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
61. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
62. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
63. Saxena, S., Herrmann, C., Hur, J., Kar, A., Norouzi, M., Sun, D., Fleet, D.J.: The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. Advances in Neural Information Processing Systems **36**, 39443–39469 (2023)
64. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
65. Shan, X., Wu, D., Zhu, G., Shao, Y., Sang, N., Gao, C.: Open-vocabulary semantic segmentation with image embedding balancing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28412–28421 (2024)
66. Shim, J.h., Yu, H., Kong, K., Kang, S.J.: Feedformer: Revisiting transformer decoder for efficient semantic segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 2263–2271 (2023)
67. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
68. Song, Y., Ermon, S.: Improved techniques for training score-based generative models. Advances in neural information processing systems **33**, 12438–12448 (2020)
69. Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)
70. Su, Y., Zhan, X., Fang, H., Li, Y.L., Lu, C., Yang, L.: Motion before action: Diffusing object motion as manipulation condition. IEEE Robotics and Automation Letters (2025)
71. Su, Y., Zhang, C., Chen, S., Tan, L., Tang, Y., Wang, J., Liu, X.: Dspv2: Improved dense policy for effective and generalizable whole-body mobile manipulation. arXiv preprint arXiv:2509.16063 (2025)
72. Tao, J., Chen, Z., Sun, Z., Guo, H., Leng, B., Yu, Z., Wang, Y., He, Z., Lei, X., Yang, J.: Seg-road: a segmentation network for road extraction based on transformer and cnn with connectivity structures. Remote Sensing **15**(6), 1602 (2023)
73. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3554–3563 (2024)
74. Tosi, F., Ramirez, P.Z., Poggi, M.: Diffusion models for monocular depth estimation: Overcoming challenging conditions. In: European Conference on Computer Vision. pp. 236–257. Springer (2024)

75. Tu, Z., Ding, Z., Wang, J.: Open-vocabulary uni-versal image segmentation with maskclip. In: ICML (2023)
76. Van Gansbeke, W., De Brabandere, B.: A simple latent diffusion approach for panoptic segmentation and mask inpainting. In: European Conference on Computer Vision. pp. 78–97. Springer (2024)
77. Wang, H., Zhou, K., Gu, B., Feng, Z., Wang, W., Sun, P., Xiao, Y., Zhang, J., Dong, H.: Transdiff: Diffusion-based method for manipulating transparent objects using a single rgb-d image. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 7277–7283. IEEE (2025)
78. Wang, J., Li, X., Zhang, J., Xu, Q., Zhou, Q., Yu, Q., Sheng, L., Xu, D.: Diffusion model is secretly a training-free open vocabulary semantic segmenter. IEEE Transactions on Image Processing (2025)
79. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. arXiv preprint arXiv:2110.08733 (2021)
80. Wang, X., Girdhar, R., Yu, S.X., Misra, I.: Cut and learn for unsupervised object detection and instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3124–3134 (2023)
81. Wang, X., Si, C., Yang, X., Zhao, Y., Wang, W., Yang, X., Shen, W.: Opmapper: Enhancing open-vocabulary semantic segmentation with multi-guidance information. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
82. Wang, Y., Yang, L., Liu, X., Yan, P.: An improved semantic segmentation algorithm for high-resolution remote sensing images based on deeplabv3+. Scientific reports **14**(1), 9716 (2024)
83. Weyler, J., Magistri, F., Marks, E., Chong, Y.L., Sodano, M., Roggiolani, G., Chebrolu, N., Stachniss, C., Behley, J.: Phenobench: A large dataset and benchmarks for semantic image interpretation in the agricultural domain. IEEE transactions on pattern analysis and machine intelligence **46**(12), 9583–9594 (2024)
84. Whitaker, J.: Multi-resolution noise for diffusion model training. Report on Weights & Biases (Feb 2023), `https://wandb.ai/johnowhitaker/multires_noise/reports/Multi-Resolution-Noise-for-Diffusion-Model-Training--VmlldzozNjYyOTU2`, last edited May 8 2023
85. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1206–1217 (2023)
86. Xian, Y., Choudhury, S., He, Y., Schiele, B., Akata, Z.: Semantic projection network for zero-and few-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8256–8265 (2019)
87. Xie, B., Cao, J., Xie, J., Khan, F.S., Pang, Y.: Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3426–3436 (2024)
88. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
89. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18134–18144 (2022)

90. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2955–2966 (2023)
91. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2945–2954 (2023)
92. Xu, M., Zhang, Z., Wei, F., Lin, Y., Cao, Y., Hu, H., Bai, X.: A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: European Conference on Computer Vision. pp. 736–753. Springer (2022)
93. Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., Liu, F.: Advances in medical image segmentation: A comprehensive review of traditional, deep learning and hybrid approaches. Bioengineering **11**(10), 1034 (2024)
94. Yan, H., Wu, M., Zhang, C.: Multi-scale representations by varying window attention for semantic segmentation. arXiv preprint arXiv:2404.16573 (2024)
95. Yao, W., Bai, J., Liao, W., Chen, Y., Liu, M., Xie, Y.: From cnn to transformer: A review of medical image segmentation models. Journal of Imaging Informatics in Medicine **37**(4), 1529–1547 (2024)
96. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
97. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14393–14402 (2021)
98. Zhang, S.C., Li, Y., Wu, Y.H., Hou, Q., Cheng, M.M.: Revisiting efficient semantic segmentation: Learning offsets for better spatial and class feature alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22361–22371 (2025)
99. Zhao, C., Sun, Y., Liu, M., Zheng, H., Zhu, M., Zhao, Z., Chen, H., He, T., Shen, C.: Diception: A generalist diffusion model for visual perceptual tasks. arXiv preprint arXiv:2502.17157 (2025)
100. Zhao, H., Puig, X., Zhou, B., Fidler, S., Torralba, A.: Open vocabulary scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2002–2010 (2017)
101. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)
102. Zhao, Z., Li, X., Shi, L., Imanpour, N., Wang, S.: Dpseg: Dual-prompt cost volume learning for open-vocabulary semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25346–25356 (2025)
103. Zheng, P., Jiang, J., Zhang, Y., Zeng, C., Qin, C., Li, Z.: Cgc-net: A context-guided constrained network for remote-sensing image super resolution. Remote Sensing **15**(12), 3171 (2023)
104. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. International Journal of Computer Vision **127**(3), 302–321 (2019)
105. Zhou, G., Chen, W., Gui, Q., Li, X., Wang, L.: Split depth-wise separable graph-convolution network for road extraction in complex environments from high-resolution remote-sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2021)

106. Zhou, L., Zhang, C., Wu, M.: D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 182–186 (2018)
107. Zhou, T., Xia, W., Zhang, F., Chang, B., Wang, W., Yuan, Y., Konukoglu, E., Cremers, D.: Image segmentation in foundation model era: A survey. arXiv preprint arXiv:2408.12957 (2024)
108. Zhu, M., Liu, Y., Luo, Z., Jing, C., Chen, H., Xu, G., Wang, X., Shen, C.: Unleashing the potential of the diffusion model in few-shot semantic segmentation. Advances in Neural Information Processing Systems **37**, 42672–42695 (2024)