SeeDiff: Off-the-Shelf Seeded Mask Generation from Diffusion Models

Joon Hyun Park[1], **Kumju Jo**[1], **Sungyong Baik**[1, 2†]

[1] Dept. of Artificial Intelligence, Hanyang University, South Korea
[2] Dept. of Data Science, Hanyang University, South Korea
{pjay258, juice0630, dsybaik}@hanyang.ac.kr

Abstract

Entrusted with the goal of pixel-level object classification, the semantic segmentation networks entail the laborious preparation of pixel-level annotation masks. To obtain pixel-level annotation masks for a given class without human efforts, recent few works have proposed to generate pairs of images and annotation masks by employing image and text relationships modeled by text-to-image generative models, especially Stable Diffusion. However, these works do not fully exploit the capability of text-guided Diffusion models and thus require a pre-trained segmentation network, careful text prompt tuning, or the training of a segmentation network to generate final annotation masks. In this work, we take a closer look at attention mechanisms of Stable Diffusion, from which we draw connections with classical seeded segmentation approaches. In particular, we show that cross-attention alone provides very coarse object localization, which however can provide initial seeds. Then, akin to region expansion in seeded segmentation, we utilize the semantic-correspondence-modeling capability of self-attention to iteratively spread the attention to the whole class from the seeds using multi-scale self-attention maps. We also observe that a simple-text-guided synthetic image often has a uniform background, which is easier to find correspondences, compared to complex-structured objects. Thus, we further refine a mask using a more accurate background mask. Our proposed method, dubbed **SeeDiff**, generates high-quality masks off-the-shelf from Stable Diffusion, without additional training procedure, prompt tuning, or a pre-trained segmentation network.

Code — https://github.com/BAIKLAB/SeeDiff.git

Introduction

As one of fundamental tasks in computer vision, semantic segmentation has a broad range of applications from medical imaging (Dolz, Desrosiers, and Ayed 2018; Oktay et al. 2018) to tracking (Yang et al. 2023). The applicability and practicability of semantic segmentation have increased with the emergence of neural networks for segmentation (Long, Shelhamer, and Darrell 2015; Ronneberger, Fischer, and Brox 2015). However, the training of semantic segmentation networks is data-hungry in nature, entailing a large amount of fine-grained pixel-level annotation and even more so for segmentation foundation models (1.1 billion masks across 11 million images). Such data-hunger nature can make the training of segmentation networks laborious, thereby limiting its practicability.

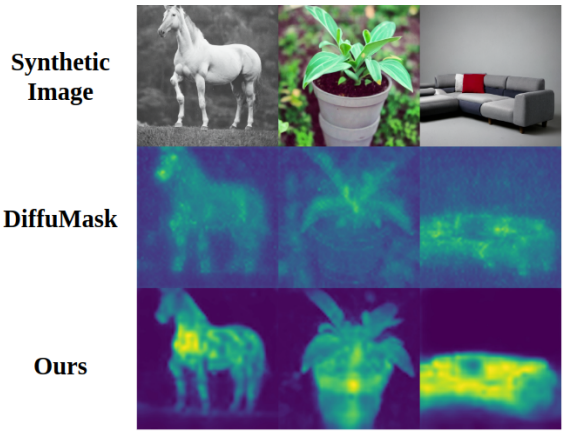

Figure 1: Comparisons against our baseline, DiffuMask (Wu et al. 2023b) on the quality of a generated mask. To better visualize the quality and accuracy of a generated mask, we visualize a soft mask before discretization. We can see that our proposed method SeeDiff produces more accurate masks with sharp boundaries, compared to our baseline.

To reduce the annotation cost, several works have employed weakly supervised learning frameworks, where only coarse labels (e.g., bounding boxes (Hsu et al. 2019), lines or points (Lin et al. 2016), or image-level class labels (Ahn, Cho, and Kwak 2019), etc.) are available. However, relying on such weak labels has led to task-specific complex designs (Khoreva et al. 2016) and relatively low performance.

Alternatively, there has been efforts on generating synthetic segmentation dataset consisting of image-mask pairs, in order to remove human efforts for annotation. Early works (Zhang et al. 2021; Li et al. 2022) have employed the features of generative adversarial networks (GANs) to generate synthetic data. However, these methods introduce ad-hoc decoders that still require a few pixel-level annotations for training.

Following the advances in generative models (Ho, Jain,

†Corresponding author.

and Abbeel 2020) with the emergence of diffusion models, few recent works (Wu et al. 2023b; Nguyen et al. 2023) have shifted to using diffusion models to generate synthetic segmentation datasets. They employ text-conditioned diffusion models (e.g., Stable Diffusion (Rombach et al. 2022)), which can generate realistic images corresponding to input texts and thereby allowing them to easily control image generation. In particular, previous works have mainly focused on image-text relationships learned by cross-attention layers. Such multimodal latent space allows them to localize which image regions correspond to particular texts/classes, resulting in mask generation. However, cross-attention maps, which the previous works heavily rely on, can provide very coarse masks. These coarse masks result in prior works resorting to the utilization of a pre-trained segmentation network trained with pixel-level annotations (Wu et al. 2023b), careful text prompt tuning (Nguyen et al. 2023), or extra training procedures (Wu et al. 2023b; Nguyen et al. 2023), which can result in poor quality masks for new classes that are unknown during the pre-training or additional training of segmentation networks and text prompt tuning.

In this work, we perform an analysis on attention maps produced by Stable Diffusion, in order to obtain high-quality masks. Upon analysis, we claim that both cross-attention layers and self-attention layers play crucial yet complementary roles. While cross-attention layers provide crude attention maps, cross-attention maps can serve as good starting points for localizing objects of interests corresponding to provided text. On the other hand, self-attention layers are well-known for modeling correspondences, which can be used to find other parts of the object that cross-attention might have missed. We also note different characteristics of self-attention across different layers: deeper layers provide low-resolution attention maps with better semantic information, while earlier layers give high-resolution fine-grained attention maps with less semantic information, corroborating previous findings (Zeiler and Fergus 2013; Dosovitskiy et al. 2021). We also observe that self-attention layers struggle to find correspondences between all parts of the whole object with non-homogeneous appearances, whereas self-attention layers find more accurate correspondences within background, which is usually uniform in a simple-text-guided synthetic image.

Founded upon our observations from analysis, we draw connections between the observed characteristics of attention layers and classical seeded segmentation algorithms (Zhang, Zheng, and Cai 2010). Seeded segmentation first finds seeds (initial starting pixels) and expand segmentation masks to pixels that are similar to seeds. With this realization, we propose a new off-the-shelf framework named **SeeDiff** that draws inspiration from seeded segmentation (Zhang, Zheng, and Cai 2010) to utilize attention maps of Stable Diffusion to generate masks. In particular, we localize a target class object that corresponds to a text, utilizing the cross-attention maps, which quantify text-image correspondences. The coarse responses of the cross-attention maps provide initial seeds, which are then used to find other parts of the object. Such mask region expansion is performed by using self-attention maps that correspond to seeds to find other missing parts of the object. We perform mask region expansion iteratively, as the feature resolution increases, to exploit semantic information from low-resolution features and fine-grained details from high-resolution features. We further refine a mask, using a background mask that can be obtained by mask region expansion from background seeds extracted from the inverted mask.

Experimental results demonstrate outstanding performance of segmentation networks trained with our generated image-mask pairs. The results underline the effectiveness of our proposed approach in generating high-quality fine-grained pixel-level annotations, without the need for a pre-trained segmentation network, text prompt tuning, training a new module, or learning procedures.

Related Works

Semantic Segmentation. Providing fine-grained yet essential information, semantic segmentation has gained attention not only from high-level computer vision tasks (e.g., scene understanding (Xiao et al. 2018)) but a broad range of other domains, such as self-driving cars (Teichmann et al. 2016) and medical image analysis (Dolz, Desrosiers, and Ayed 2018; Oktay et al. 2018). Since the emergence of deep neural networks (DNNs), the performance has elevated with advances in neural network architecture designs (Long, Shelhamer, and Darrell 2015; Yu and Koltun 2016; Chen et al. 2016; Ronneberger, Fischer, and Brox 2015; Noh, Hong, and Han 2015; Badrinarayanan, Kendall, and Cipolla 2015; Sharma et al. 2020; Kipf and Welling 2017), especially with transformers (Xie et al. 2021; Zheng et al. 2021; Cheng, Schwing, and Kirillov 2021; Cheng et al. 2022) and foundation models (Kirillov et al. 2023; Zou et al. 2023). However, behind the exceptional performance of DNN-based segmentation models lies a significant amount of human efforts in collecting and annotating large-scale data required for training DNNs. This is especially the case for transformer-based networks and foundation models, which are known for its data hunger (Kirillov et al. 2023). To make things worse, segmentation networks require fine-grained pixel-level annotations, making dataset preparation for training even more laborious. To reduce the annotation cost of semantic segmentation, several works have proposed to make use of cheaper annotations, such as image-wise category labels (Hsu et al. 2019), points/lines (Lin et al. 2016), and bounding boxes (Ahn, Cho, and Kwak 2019). However, these weak labels lead to performance degradation. As such, recent models, such as SAM and Mask2Former, still use pixel-level annotations to maximize performance.

Dataset Generation for Semantic Segmentation. To minimize human efforts in collecting annotation, a line of works have tried to borrow the data synthesis capability of generative models to generate synthetic datasets consisting of image-mask pairs. Early methods (Voynov, Morozov, and Babenko 2020; Melas-Kyriazi et al. 2022; Zhang et al. 2021; Li et al. 2022) employ generative adversarial networks (GANs) (Goodfellow et al. 2014), where semantic features from intermediate layers are extracted and passed

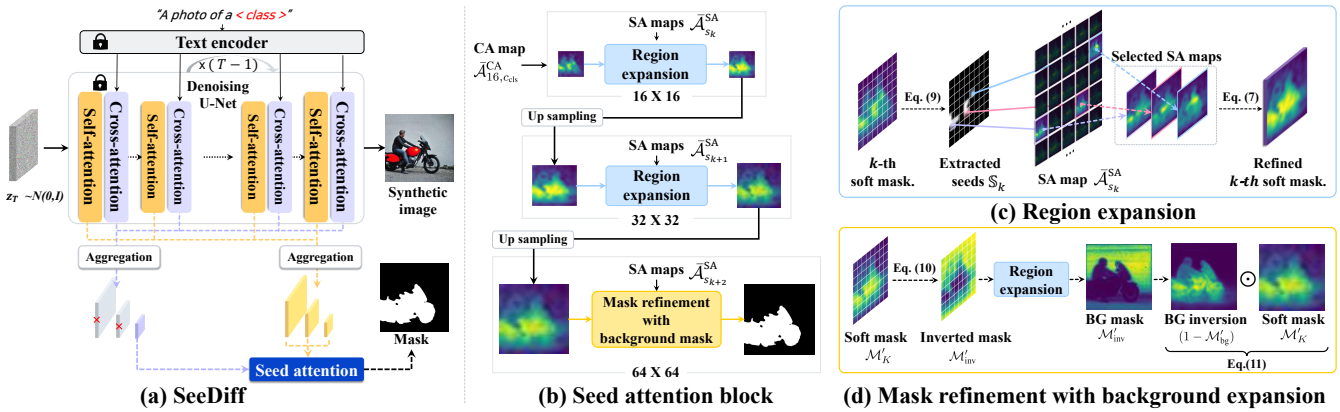

Figure 2: Overall framework: (a) Our method SeeDiff is an off-the-shelf mechanisms that generates a mask by utilizing extract attention maps from a pre-trained Stable Diffusion conditioned on an input prompt. (b) Our mask generation draws inspiration from a classical seeded segmentation process, where a cross-attention (CA) map provides initial cues to the target object location, which are then iteratively expanded to object regions through (c) region expansion with SA maps. (d) A mask is further refined with a background mask, which is similarly obtained via seed extraction and region expansion of an inverted mask.

to need-to-train decoders for mask generation. However, these works either require optimization process and can only perform foreground-background separation (Voynov, Morozov, and Babenko 2020; Melas-Kyriazi et al. 2022) or extra decoders (Zhang et al. 2021; Li et al. 2022) that still need to be trained with few images annotated with pixel-level fine-grained masks. Recent works (Li et al. 2023; Wu et al. 2023b,a; Nguyen et al. 2023) have shifted to diffusion-based (Ho, Jain, and Abbeel 2020) text-to-image generative models (e.g., GLIDE (Nichol et al. 2022), Imagen (Saharia et al. 2022), and Stable Diffusion (Rombach et al. 2022)) for its breakthrough in generation performance and easy control over generation via text. While using diffusion models has led to high-quality image-mask generation, they require either a pre-trained object detector (Li et al. 2023) or a pre-trained segmentation network (Wu et al. 2023b); extra modules that need to be trained (Li et al. 2023; Wu et al. 2023b); or extra learning processes (Nguyen et al. 2023; Ma et al. 2023) for refinement. These processes can lead to poor quality mask generation for classes unknown during such processes, thereby being unable to fully exploit the capabilities of text-to-image generative models trained on a very large-scale dataset. Our work, on the other hand, draws inspiration from seeded segmentation, which we claim to be easily transferable to image-mask generation task with off-the-shelf text-guided diffusion models. As a result, we synthesize high-quality image-mask datasets, without the need for pre-trained segmentation/detection networks, extra trainable modules, or optimization/learning processes.

Background

Problem Formulation

The objective of this work is to construct a high-quality synthetic segmentation dataset $\mathcal{D} = \{\boldsymbol{x}_n, \boldsymbol{m}_n\}_{n=1}^{N}$ consisting of synthetic images $\boldsymbol{x}_n$ with corresponding pixel-level masks $\boldsymbol{m}_n$. In order to have correspondences between generated images and masks, we first need to have control over which objects will be generated in images. To this end, similar to previous works (Wu et al. 2023b; Nguyen et al. 2023), we generate images via a text-to-image generative model, which allows for a control over image generation via texts. Following prior works (Wu et al. 2023b; Nguyen et al. 2023), we use instantiate a text-to-image generative model with Stable Diffusion (Rombach et al. 2022), which is one of recent, successful, and open-source text-to-image generative models.

Stable Diffusion

Stable Diffusion (Rombach et al. 2022) is a text-guided diffusion model that aims to generate an image $\boldsymbol{x}$ from a random noise $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, guided by a text embedding $\boldsymbol{\tau}(y)$ with an input text prompt y of length P and a text encoder $\boldsymbol{\tau}$. Diffusion models assume that a random noise $\boldsymbol{z}$ and $\boldsymbol{x}$ are connected by Markov chain that gradually adds noise to $\boldsymbol{x} = \boldsymbol{z}_0$ to form $\boldsymbol{z} = \boldsymbol{z}_T$ over a pre-defined number of time steps T:

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{t-1}) = \mathcal{N}(\boldsymbol{z}_t; \sqrt{1-\beta_t}\boldsymbol{z}_{t-1}, \beta_t\mathbf{I}), \quad (1)$$

where β_t is a variance or noise schedule. Under the assumption, diffusion models formulate image generation as a progressive noise removal (i.e., denoising) from a noisy image to obtain gradually less noisy image and eventually a clean image at the end. Hence, a denoising network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$, parameterized by parameters $\boldsymbol{\theta}$, is trained to estimate an added noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ guided by conditioning text prompt y at each time step t, optimizing the following objective function:

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t,\boldsymbol{\tau}(y)}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, \boldsymbol{\tau}(y))\right\|_2^2\right]. \quad (2)$$

Then, during image generation process, a noise estimated by a denoising network is removed from a noisy image to gradually obtain a less noisy image, leading to a new clean generated image after a pre-defined number of steps T. In contrast to previous diffusion models, Stable Diffusion performs

noise addition and denoising process on the latent space, instead of pixel space.

A denoising network in Stable Diffusion employs U-Net (Ronneberger, Fischer, and Brox 2015) architecture consisting of residual blocks and transformer blocks, each of which is composed of a cross-attention (CA) and self-attention (SA) layer. There are $L = 16$ transformer blocks in total. At each l-th transformer layer, a l-th CA layer fuses information from the embedding of a conditioning text prompt $\boldsymbol{\tau}(y) \in \mathbb{R}^{P \times d_\tau}$ and latent features $\boldsymbol{z}_t^l \in \mathbb{R}^{H_l \times W_l \times d_z}$ from a previous layer, aiding in generating images that correspond to a conditioning text prompt. During this process, a l-th CA layer produces a feature map (CA map) $\mathcal{A}_{l,t}^{\text{CA}} \in \mathbb{R}^{H_l \times W_l \times P}$ at each time step t as follows:

$$\mathcal{A}_{l,t}^{\text{CA}} = \text{softmax}\left(\frac{Q_{l,t}^z \cdot {K_l^\tau}^\top}{\sqrt{d_l}}\right), \tag{3}$$

where $Q_{l,t}^z, K_l^\tau$ are a query matrix of $\boldsymbol{z}_t^l$ and a key matrix of $\boldsymbol{\tau}(y)$ obtained via learned linear projections, respectively; and d_l is the dimension of latent features at l-th layer. On the other hand, a l-th SA layer is used to learn spatial similarities and correspondences within latent features $\boldsymbol{z}_t^l$ from a previous layer. Similarly, self-attention process produces a SA map $\mathcal{A}_{l,t}^{\text{SA}} \in \mathbb{R}^{H_l \times W_l \times H_l \times W_l}$ via

$$\mathcal{A}_{l,t}^{\text{SA}} = \text{softmax}\left(\frac{Q_{l,t}^z \cdot {K_{l,t}^z}^\top}{\sqrt{d_l}}\right), \tag{4}$$

where $Q_{l,t}^z, K_{l,t}^z$ are query and key matrices of $\boldsymbol{z}_t^l$ via learned linear projections.

Since Stable Diffusion employs U-Net, attention layers in encoders and decoders produce attention maps of different resolutions. Specifically, attention maps have four different resolutions in total: $(H_l, W_l) \in \{(s_r, s_r)\}_{r=0}^3$, where $s_r \in \{8, 16, 32, 64\}$. While there are numerous attention maps generated across all diffusion time steps $t \in [1, T]$ and layers $l \in [1, 16]$, the fact that there are only four different resolutions suggests that attention maps can be grouped together according to resolution.

Aggregation of Attention Maps

To summarize information from a myriad of generated attention maps, attention maps can be grouped according to resolution and then aggregated via averaging and normalization to have values between 0 and 1 (Wu et al. 2023b; Nguyen et al. 2023):

$$\bar{\mathcal{A}}_{s_r} = \frac{1}{|\mathbb{L}_{s_r}| \cdot T} \sum_{l \in \mathbb{L}_{s_r}} \sum_{1 \le t \le T} \frac{\mathcal{A}_{l,t}}{\max(\mathcal{A}_{l,t})}, \tag{5}$$

where $\mathbb{L}_{s_r}$ is the set of indices of layers that produce attention maps of resolution with scale s_r; $\bar{\mathcal{A}}_{s_r}$ is an aggregated attention map of resolution with scale s_r; and $\mathcal{A}_{l,t}$ is either CA map or SA map. Thus, $\bar{\mathcal{A}}_{s_r}^{\text{SA}}$ and $\bar{\mathcal{A}}_{s_r}^{\text{CA}}$ represents an aggregated SA map and aggregated CA map of resolution s_r, respectively. Previous works have solely used $\bar{\mathcal{A}}_{s_r}^{\text{CA}}$ or its simple multiplication with $\bar{\mathcal{A}}_{s_r}^{\text{SA}}$ to find regions corresponding to text in the course of generating a mask. Hereafter, we will refer to $\bar{\mathcal{A}}_{s_k}^{\text{CA}}$ and $\bar{\mathcal{A}}_{s_k}^{\text{SA}}$ as CA maps and SA maps, respectively.

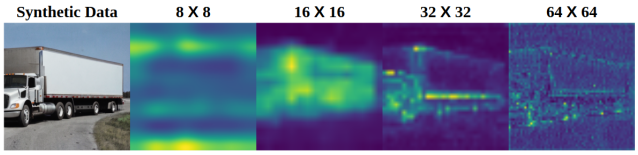

Figure 3: Cross-attention (CA) map at different scales.

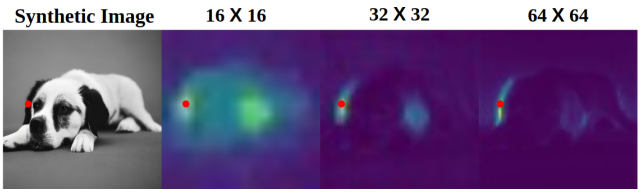

Figure 4: Self-attention (SA) maps at different scales. Red dot indicates the coordinates the visualized multiscale SA maps correspond to.

Proposed Method

In this work, we aim to extract different yet complementary information out of both cross-attention (CA) and self-attention (SA) maps for the purpose of generating high-quality masks corresponding to generated images and text. We use CA maps as initial cues (i.e., seeds) to the location of the object at seed intialization. Seeds, in turn, are then used to expand regions by selecting SA maps that highlight regions similar to the seeds at region expansion process. To capture fine-grained details, we first attend to whole object and then fine-grained details by iteratively extracting seed from a mask from a previous step (lower resolution) and expanding regions using SA maps of higher resolution. We further refine a mask by toning down the attention on background and strengthening attention on foreground by using background masks obtained by extracting background seeds expanded to the whole background via SA region expansion, as delineated at mask refinement using background mask process.

Seed Initialization with Cross-Attention Map

We start with reiterating that CA maps quantify how similar latent features $\boldsymbol{z}$ are to a text embedding $\tau(y)$. In particular, after simplifying a notation without loss of generality, $\bar{\mathcal{A}}^{\text{CA}}[i, j, c]$ quantifies the similarity between a latent vector $\boldsymbol{z}$ at (i, j) (i.e., $\boldsymbol{z}[i, j]$) and the embedding of c-th token in a text prompt y (i.e., $\boldsymbol{\tau}(y)[c]$). Thus, we can obtain a CA map that highlights regions that correspond to a class of interest via $\bar{\mathcal{A}}^{\text{CA}}[:, :, c_{\text{cls}}]$ (or $\bar{\mathcal{A}}_{c_{\text{cls}}}^{\text{CA}}$), where c_{cls} is the index of a class token in the text prompt. Indeed, Fig. 3 shows how a CA map $\bar{\mathcal{A}}_{s_r, c_{\text{cls}}}^{\text{CA}}$ focuses on an object of interest across all scales. However, the figure exhibits different characteristics across resolution scales. While a CA map becomes more fine-grained as resolution increases, a CA map tends to attend more uniformly across the whole image and focus less

on the object. On the other hand, a lower-resolution CA map may focus better on the object but provides very coarse semantic correspondence information, losing a lot of details on the structure of the object. In order to strike a balance between details and precision, we select a 16×16 CA map (i.e., $\bar{\mathcal{A}}^{\text{CA}}_{16,c_{\text{cls}}}$) to localize the object of interest. However, we observe that a 16×16 CA map still emphasizes only few specific parts of an object and lacks sharp boundaries, making it inappropriate to be used as final mask. Upon observation, we propose to obtain initial cues (i.e., seeds) as to the location of an object from a 16×16 CA map by collecting spatial coordinates at which a CA value is larger than a threshold α as follows:

$$\mathbb{S}_1 = \{(i,j) \mid \bar{\mathcal{A}}^{\text{CA}}_{16,c_{\text{cls}}}[i,j] \geq \alpha\}. \tag{6}$$

Iterative Seed Extraction and Region Expansion with Self-Attention Maps

While the obtained seeds may locate the object, these seeds attend to only a small portion of it. Hinged on the seeds, to expand the attention to the whole object, we turn our attention to the effectiveness of self-attention in learning the similarities between features. In particular, we note that multiscale SA maps have different levels of attention details at each scale. Fig. 4 demonstrates that lower-resolution SA maps attend to broad aspects of the object, while higher-resolution SA maps attend to fine-grained parts of the object. This observation motivates us to formulate an iterative region expansion, where we first expand to the broad aspects and then fine-grained aspects of the objects.

Formally, with a simplified notation without loss of generality, $\bar{\mathcal{A}}^{\text{SA}}[i,j,:,:]$ quantifies the correlation of features $\boldsymbol{z}$ between (i,j) and all locations. This means that we can attend to other parts of the object by simply indexing $\bar{\mathcal{A}}^{\text{SA}}$ with coordinates of seeds:

$$\mathcal{M}'_k = \frac{1}{|\mathbb{S}_k|} \sum_{(i,j)\in\mathbb{S}_k} \bar{\mathcal{A}}^{\text{SA}}_{s_k}[i,j,:,:]. \tag{7}$$

Then, we upsample an obtained mask to a higher resolution s_{k+1}, such that we can perform a similar process with SA maps of higher resolution $\bar{\mathcal{A}}^{\text{SA}}_{s_{k+1}}$:

$$\mathcal{M}_{k+1} = \text{bilinear-upsample}_{s_{k+1}}(\mathcal{M}'_k), \tag{8}$$

where $\text{bilinear-upsample}_{s_{k+1}}$ performs upsampling to s_{k+1} resolution via bilinear-interpolation.

Since we have already located the object, we do not utilize CA maps at higher resolution, which has been shown to give imprecise attention (i.e., attend to background). Instead, we extract seeds from $\mathcal{M}_{k+1}$:

$$\mathbb{S}_{k+1} = \{(i,j) \mid \mathcal{M}_{k+1}[i,j] \geq \alpha\}. \tag{9}$$

Then, we repeat the process of a mask expansion and refinement using Eq. 7, 8, 9 until the last iteration K, where the highest resolution of SA maps is reached (i.e., $s_K = 64$). In the last iteration, we only perform region expansion and do not perform upsampling and seed extraction.

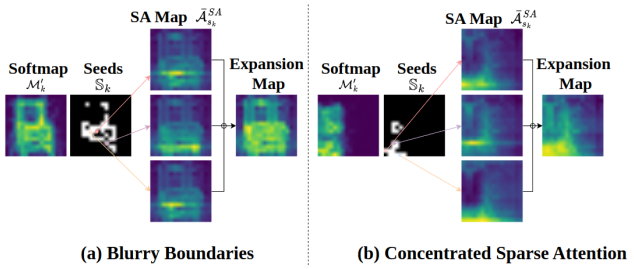

Figure 5: Persistent errors from CA map seeds. (a) CA map seeds can point to background region. (b) CA map seeds can be sparse and concentrated.

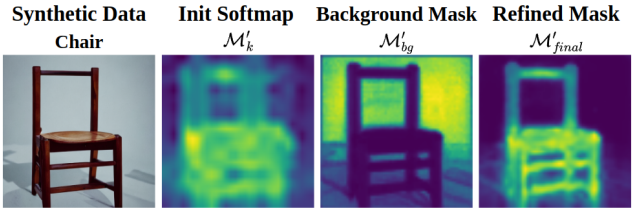

Figure 6: Mask refinement with a background mask.

Mask Refinement with Background Mask

While the iterative region expansion process with SA Maps has substantially enhanced the quality of a mask, we observe that initial seeds from CA maps introduce two persistent errors that propagate throughout the process, as illustrated in Fig. 5. The first error is that the initial seeds sometimes point at background region near the object boundary, especially when objects have detailed and complex structures. These inaccurate seeds result in region expansion to nearby background regions, thereby creating blurry boundary after region expansion with SA, as depicted in Fig. 5(a). The second error is that seeds are often sparse and concentrated at the certain parts of an object with non-homogeneous appearances, as CA finds strong correlation between certain parts of an object and a text. When seeds are sparse and concentrated and features of objects have large variances, region expansion with SA maps fails to resolve the issue, as shown in Fig. 5(b).

To resolve these two issues, we start with an observation that a synthetic image generated by Stable Diffusion with a simple text prompt generates a relatively uniform background. Upon our observation, we claim that it is much easier for SA maps to give uniformly distributed attention to background. This means we can refine a mask of class by down-weighting background parts and up-weighting class regions, via subtraction of an accurate background mask from a mask of class. To obtain a background mask, we simply inverse a mask obtained from the last iteration of the region expansion process in region expansion:

$$\mathcal{M}'_{\text{inv}} = (1 - \mathcal{M}'_K), \tag{10}$$

from which background seeds $\mathbb{S}_{\text{bg}}$ are obtained, similar to Eq. 9. Then, a background mask $\mathcal{M}'_{\text{bg}}$ is finally obtained from $\mathbb{S}_{\text{bg}}$ via region expansion with SA maps (Eq. 7). Then,

Train Set	Number	Backbone	Prompt Tuning LLM	Prompt Tuning Retrieval	Semantic Segmentation (IoU) for Selected Classes (%) aeroplane	bus	cat	chair	cow	dog	horse	person	sheep	sofa	mIoU (%) val
Training with Pure Real Data															
VOC	R: 10.6k (all)	ResNet50	✗	✗	87.5	95.5	92.2	44.0	85.4	89.1	82.1	89.2	80.6	53.6	77.3
	R: 10.6k (all)	Swin-B	✗	✗	97.0	91.7	96.5	57.5	95.9	96.8	94.4	92.5	95.1	65.6	84.3
	R: 5.0k	Swin-B	✗	✗	95.5	96.1	95.2	47.3	90.3	92.8	94.6	90.9	93.7	61.4	83.4
*Training with **Pure Synthetic Data***															
Dataset Diffusion	S: 40.0k	ResNet50	✓	✓	-	-	-	-	-	-	-	-	-	-	60.4
DiffuMask	S: 60.0k	ResNet50	✗	✓	80.7	81.2	79.3	14.7	63.4	65.1	64.6	71.0	64.7	27.8	57.4
	S: 60.0k	Swin-B	✗	✓	90.8	88.3	92.5	27.2	92.2	86	89	76.5	92.2	49.8	70.6
SeeDiff (Ours)	S: 40.0k	ResNet50	✗	✗	81.6	75.8	81.8	13.2	59.0	62.5	59.3	61.6	70.2	48.5	61.2
	S: 40.0k	Swin-B	✗	✗	94.8	93.1	88.0	22.9	90.9	87.3	71.9	73.0	94.7	63.1	78.6
	S: 60.0k	ResNet50	✗	✗	82.3	75.6	82.7	15.6	62.2	61.7	64.1	67.5	73.0	51.2	62.6
	S: 60.0k	Swin-B	✗	✗	95.5	92.0	91.5	22.4	91.5	85.1	78.3	74.9	94.2	65.3	**79.6**
*Finetune on **Real Data***															
DiffuMask	S: 60.0k+R: 5.0k	ResNet50	✗	✓	85.4	92.9	91.7	38.4	86.5	86.2	82.5	87.5	81.2	39.8	77.6
	S: 60.0k+R: 5.0k	Swin-B	✗	✓	95.6	96.9	96.6	51.5	96.7	95.5	96.1	91.5	96.4	70.2	84.9
SeeDiff (Ours)	S: 60.0k+R: 5.0k	ResNet50	✗	✗	91.9	82.0	93.7	25.7	67.9	77.0	74.6	76.6	73.8	56.0	71.4
	S: 60.0k+R: 5.0k	Swin-B	✗	✗	98.5	98.6	96.7	49.9	89.5	91.8	98.4	92.0	97.4	67.8	**88.8**

Table 1: Semantic segmentation results on VOC 2012 `val`. 'S' and 'R' denote synthetic images and real images, respectively.

we obtain a more refined mask by multiplying a class-region mask $\mathcal{M}'_K$ with the inverse of a background mask $\mathcal{M}'_{\text{bg}}$, such that background regions are diminished while class regions are amplified:

$$\mathcal{M}'_{\text{final}} = (1 - \mathcal{M}'_{\text{bg}}) \odot \mathcal{M}'_K, \tag{11}$$

which is then upsampled to full resolution 512×512 to produce a higher resolution mask $\mathcal{M}_{\text{final}}$, followed by binarization with a threshold β to obtain a final annotation mask $\boldsymbol{m}$ for a synthetic image $\boldsymbol{x}$:

$$\boldsymbol{m}[i,j] = \mathcal{M}_{\text{final}}[i,j] \odot \mathbb{1}_{\{\mathcal{M}_{\text{final}}[i,j] \geq \beta\}}, \tag{12}$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function which gives 1 when the condition in subscript is met and 0 otherwise.

Experiments

Experimental Settings

Datasets. Following the settings of the previous work DiffuMask (Wu et al. 2023b), we evaluated our model on the following two datasets Pascal-VOC2012 (Everingham et al. 2010) and Cityscapes (Cordts et al. 2016). In particular, we also conducted an evaluation on the Pascal-VOC2012 dataset in the open-vocabulary semantic segmentation setting, where the evaluation is conducted by dividing into 15 seen classes and 5 unseen classes, following the setup of the previous model (Ding et al. 2022; Wu et al. 2023b).

Evaluation Metric. We use mIoU score as a metric to evaluate the segmentation task, which calculates the intersection between the predicted regions and the actual ground truth masks, and then takes the average value.

Implementation Details. We construct our components using a pre-trained Stable Diffusion (Rombach et al. 2022) model. We utilize the Stable Diffusion 2-base version to generate images with $T = 50$ timesteps as denoising step. We use only the pre-trained internal features without any fine-tuning of the Stable Diffusion model. As for text prompts, we generate the dataset using the widely used text prompt in the form of 'a photo of a $<$ `CLASS` $>$'. We utilize $\alpha = 0.5$ as a threshold parameter to extract the seeds and $\beta = 0.3$ as a threshold parameter to discretize a soft mask to a final mask. Following previous works (Wu et al. 2023b; Nguyen et al. 2023), to evaluate the generated images and masks, we train and evaluate the segmentation model Mask2Former (Cheng et al. 2022). The settings required for training and evaluating Mask2Former, including initialization, data augmentation, batch size, weight decay, and learning rate, are configured according to the original paper. In the PascalVOC-2012 (Everingham et al. 2010) setting, we generate $2k$ and $3k$ images per class, using a total number of images ($40.0k$ and $60.0k$) identical to those used in previous studies such as Diffusion Dataset (Nguyen et al. 2023) and DiffuMask (Wu et al. 2023b). For data augmentation, we apply the same techniques used in DiffuMask, including splicing, gaussian noise, perspective transformation, and occlusion. In the Cityscapes dataset, we conduct evaluations only on the 'Human' and 'Vehicle' categories, following the same evaluation setting as same as our baseline DiffuMask (Wu et al. 2023b). The 'Human' and 'Vehicle' categories consist of six sub-classes: person, rider, car, bus, truck, and train. All experiments, including image generation and evaluation, have been conducted on NVIDIA 4090 RTX GPU. Since we utilize only the internal parameters of Stable Diffusion, there is no significant increase in computational cost.

Experimental Results

Semantic Segmentation on VOC 2012. Table 1 shows the results of semantic segmentation on the VOC 2012 dataset. We generate 40k and 60k images to match the number of generated images for a fair comparison with previous work. (Wu et al. 2023b; Nguyen et al. 2023) Training Mask2Former (Cheng et al. 2022) with our generated data results in significant performance improvement. Additional fine-tuning with real data brings further improvement, surpassing the performance of a model (Swin-B backbone)

Module				mIoU (%)
CAA	**CRE**	**IRE**	**BE**	$\text{VOC}_{\texttt{val}}$
✓	✗	✗	✗	68.8
✓	✓	✗	✗	71.0
✓	✓	✓	✗	73.2
✓	✓	✓	✓	**79.6**

Table 2: Module Ablations. We perform ablations of our module with VOC 2012 val, using Mask2Former with Swin-B. CAA denotes cross-attention aggregation, representing our baseline. CRE denotes cross-attention guided region expansion. IRE denotes iterative region expansion. BE denotes background expansion inversion.

Train Set	Number	Backbone	IoU% Human	IoU% Vehicle
Train with Pure Real Data				
Cityscapes	3.0k (all)	Swin-B	85.5	96.0
	1.5k	Swin-B	84.6	95.3
Train with Pure Synthetic Data				
DiffuMask	100.0k	Swin-B	72.1	87.0
SeeDiff (ours)	100.0k	Swin-B	74.3	89.8

Table 3: Result of Semantic Segmentation on Cityscapes val.

trained solely with the whole real data by a large margin ($4 \sim 5\%$). We also note that our generated images bring substantial performance improvements, compared to baselines, without any prompt tuning. Also, even when using less number of images, SeeDiff outperforms DiffuMask. The results underline the effectiveness of our method SeeDiff in generating high-quality masks.

Cityscapes. Table 3 displays the results of semantic segmentation on Cityscapes, where similar tendencies are observed. Again for Cityscapes dataset, we use input text with fixed templates. Considering the domain of Cityscapes, which consists of urban street scenes, we use the fixed prompt 'a photo of a $<$ CLASS $>$ in the city.'

Zero-Shot Semantic Segmentation. In the zero-shot setting of VOC 2012, our model outperform DiffuMask and several models trained with real data. Notably, we achieve similar performance on 'seen' and 'unseen' categories, while DiffuMask has a drop in performance on 'unseen categories. We believe this is because DiffuMask uses a pre-trained segmentation network, which limits the performance on classes unseen during the pre-training of a segmentation network.

Ablation Study

In this section, extensive ablation studies are conducted to assess the effectiveness of each proposed module. We perform ablation studies on Pascal-VOC2012, where 60k images are generated to train Mask2Former (Swin-B).

Resolution	mIoU (%) $\text{VOC}_{\texttt{val}}$
16	**65.9**
32	64.3
64	58.2

Table 4: Study on the influence of resolution of CA map. We evaluate the influence of resolution of CA map on the final performance.

Method	mIoU (%) $\text{VOC}_{\texttt{val}}$
CA	68.8
CA·SA	72.0
SeeDiff	**79.6**

Table 5: Comparisons of Attention-based Mask Generation Strategy. CA denotes cross-attention aggregation (DiffuMask). CA·SA denotes simple multiplication of CA and SA (Dataset Diffusion).

Module Ablations. Table 2 assesses the effectiveness of each component of our framework SeeDiff. In the case of Cross-Attention Aggregation (CAA), the final mask is generated solely from the aggregated cross-attention maps during the denoising step, similar to the mask generation method used by a baseline, DiffuMask (Wu et al. 2023b). Cross-attention guided region expansion (CRE) generates the final mask by aggregating the self-attention maps corresponding to the seeds extracted from the CA map at each resolution of the U-Net. Iterative region expansion (IRE) denotes a variant of our method, which uses a soft mask $\mathcal{M}'_k$ from low resolution as an initial mask for the next resolution. Finally, IRE with mask refinement with background expansion (BE) represents our final model.

Study on Different Feature Scales. Table 4 evaluates the influence of CA maps of different scales. Through this experiment, we observe that higher resolutions tend to lose objectness, as also exhibited in Figure 3. This validates the our design choice of using CA map of 16×16 resolution.

Study on Mask Generation Strategy. Table 5 compares different mask generation strategies: CA (Cross-Attention-based mask generation; DiffuMask (Wu et al. 2023b)), CA·SA (Cross-Attention-Self-Attention-Multiplication-based mask generation; Dataset Diffusion (Nguyen et al. 2023)), and our SeeDiff. SeeDiff demonstrates to be most effective. The results validate our motivation of drawing connections with classical seeded segmentation algorithms.

Conclusion

In this study, we introduce a novel off-the-shelf framework, named **SeeDiff**, that generates high-quality image-mask pairs from Stable Diffusion, without any training, prompt tuning, or using pre-trained segmentation networks. The design of a framework is inspired by a classical seeded segmentation algorithm, which expands initial point cues (i.e., seeds). We observe that CA maps provide coarse localization of objects, serving as seeds. In the mean time, we utilize SA map to perform mask region expansion to cover the whole object. Consequently, SeeDiff generates high-quality image-mask pairs that can be used to train a segmentation network with outstanding performance.

Acknowledgments

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)

References

Adams, R.; and Bischof, L. 1994. Seeded region growing. *Transactions on Pattern Analysis and Machine Intelligence*.

Ahn, J.; Cho, S.; and Kwak, S. 2019. Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations. In *CVPR*.

Ahn, J.; and Kwak, S. 2018. Learning Pixel-Level Semantic Affinity With Image-Level Supervision for Weakly Supervised Semantic Segmentation. In *CVPR*.

Akiva, P.; and Dana, K. 2021. Towards Single Stage Weakly Supervised Semantic Segmentation. arXiv:2106.10309.

Amit, T.; Shaharbany, T.; Nachmani, E.; and Wolf, L. 2022. SegDiff: Image Segmentation with Diffusion Probabilistic Models. arXiv:2112.00390.

Araslanov, N.; and Roth, S. 2020. Single-Stage Semantic Segmentation From Image Labels. In *CVPR*.

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2015. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*.

Baek, D.; Oh, Y.; and Ham, B. 2021. Exploiting a Joint Embedding Space for Generalized Zero-Shot Semantic Segmentation. In *ICCV*.

Baranchuk, D.; Rubachev, I.; Voynov, A.; Khrulkov, V.; and Babenko, A. 2022. Label-Efficient Semantic Segmentation with Diffusion Models. In *ICLR*.

Boykov, Y. Y.; and Jolly, M.-P. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *ICCV*.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *ICLR*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *CVPR*.

Bucher, M.; Vu, T.-H.; Cord, M.; and Pérez, P. 2019. Zero-Shot Semantic Segmentation. In *NeurIPS*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*.

Cha, J.; Mun, J.; and Roh, B. 2023. Learning to Generate Text-grounded Mask for Open-world Semantic Segmentation from Only Image-Text Pairs. In *CVPR*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K. P.; and Yuille, A. L. 2016. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *Transactions on Pattern Analysis and Machine Intelligence*.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *CVPR*.

Cheng, B.; Schwing, A. G.; and Kirillov, A. 2021. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *NeurIPS*.

Cheng, J.; Nandi, S.; Natarajan, P.; and Abd-Almageed, W. 2021. SIGN: Spatial-information Incorporated Generative Network for Generalized Zero-shot Semantic Segmentation. In *ICCV*.

Choi, J.; Kim, T.; and Kim, C. 2019. Self-Ensembling with GAN-based Data Augmentation for Domain Adaptation in Semantic Segmentation. In *ICCV*.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*.

Devaranjan, J.; Kar, A.; and Fidler, S. 2020. Meta-Sim2: Unsupervised Learning of Scene Structure for Synthetic Data Generation. In *ECCV*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling Zero-Shot Semantic Segmentation. In *CVPR*.

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. NICE: Nonlinear Independent Components Estimation. In *ICLR*.

Dolz, J.; Desrosiers, C.; and Ayed, I. B. 2018. IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet. In *MICCAI*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming Transformers for High-Resolution Image Synthesis. In *CVPR*.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Networks. In *NeurIPS*.

Grady, L. 2006. Random walks for image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*.

Gu, Z.; Chen, H.; Xu, Z.; Lan, J.; Meng, C.; and Wang, W. 2024. DiffusionInst: Diffusion Model for Instance Segmentation. In *ICASSP*.

Gu, Z.; Zhou, S.; Niu, L.; Zhao, Z.; and Zhang, L. 2020. Context-aware Feature Generation for Zero-shot Semantic Segmentation. In *ACM MM*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-Prompt Image Editing with Cross Attention Control. In *ICLR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*.

Hsu, C.-C.; Hsu, K.-J.; Tsai, C.-C.; Lin, Y.-Y.; and Chuang, Y.-Y. 2019. Weakly Supervised Instance Segmentation using the Bounding Box Tightness Prior. In *NeurIPS*.

Huang, Z.; Wang, X.; Wang, J.; Liu, W.; and Wang, J. 2018. Weakly-Supervised Semantic Segmentation Network with Deep Seeded Region Growing. In *CVPR*.

Ikonomatakis, N.; Plataniotis, K.; Zervakis, M.; and Venetsanopoulos, A. 1997. Region growing and region merging image segmentation. In *ICDSP*.

Ji, Y.; Chen, Z.; Xie, E.; Hong, L.; Liu, X.; Liu, Z.; Lu, T.; Li, Z.; and Luo, P. 2023. DDP: Diffusion Model for Dense Visual Prediction. In *ICCV*.

Kar, A.; Prakash, A.; Liu, M.-Y.; Cameracci, E.; Yuan, J.; Rusiniak, M.; Acuna, D.; Torralba, A.; and Fidler, S. 2019. Meta-Sim: Learning to Generate Synthetic Datasets. In *ICCV*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*.

Khoreva, A.; Benenson, R.; Hosang, J.; Hein, M.; and Schiele, B. 2016. Simple Does It: Weakly Supervised Instance and Semantic Segmentation. In *CVPR*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*.

Kolesnikov, A.; and Lampert, C. H. 2016. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *ECCV*.

Krähenbühl, P.; and Koltun, V. 2012. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In *NeurIPS*.

Le, M.-Q.; Nguyen, T. V.; Le, T.-N.; Do, T.-T.; Do, M. N.; and Tran, M.-T. 2024. MaskDiff: Modeling Mask Distribution with Diffusion Probabilistic Model for Few-Shot Instance Segmentation. In *AAAI*.

Li, D.; Ling, H.; Kim, S. W.; Kreis, K.; Barriuso, A.; Fidler, S.; and Torralba, A. 2022. BigDatasetGAN: Synthesizing ImageNet with Pixel-wise Annotations. In *CVPR*.

Li, Z.; Zhou, Q.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Open-vocabulary Object Segmentation with Diffusion Models. In *ICCV*.

Lin, D.; Dai, J.; Jia, J.; He, K.; and Sun, J. 2016. ScribbleSup: Scribble-Supervised Convolutional Networks for Semantic Segmentation. In *CVPR*.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. In *CVPR*.

Ma, C.; Yang, Y.; Ju, C.; Zhang, F.; Liu, J.; Wang, Y.; Zhang, Y.; and Wang, Y. 2023. DiffusionSeg: Adapting Diffusion Towards Unsupervised Object Discovery. arXiv:2303.09813.

Mehnert, A.; and Jackway, P. T. 1997. An improved seeded region growing algorithm. *Pattern Recognition Letters.*

Melas-Kyriazi, L.; Rupprecht, C.; Laina, I.; and Vedaldi, A. 2022. Finding an Unsupervised Image Segmenter in Each of Your Deep Generative Models. In *ICLR*.

Murez, Z.; Kolouri, S.; Kriegman, D.; Ramamoorthi, R.; and Kim, K. 2018. Image to Image Translation for Domain Adaptation. In *CVPR*.

Nguyen, Q.; Vu, T.; Tran, A.; and Nguyen, K. 2023. Dataset Diffusion: Diffusion-based Synthetic Dataset Generation for Pixel-Level Semantic Segmentation. In *NeurIPS*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*.

Noh, H.; Hong, S.; and Han, B. 2015. Learning Deconvolution Network for Semantic Segmentation. In *ICCV*.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; Glocker, B.; and Rueckert, D. 2018. Attention U-Net: Learning Where to Look for the Pancreas. In *MIDL*.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pastore, G.; Cermelli, F.; Xian, Y.; Mancini, M.; Akata, Z.; and Caputo, B. 2021. A Closer Look at Self-training for Zero-Label Semantic Segmentation. In *CVPRW*.

Pnvr, K.; Singh, B.; Ghosh, P.; Siddiquie, B.; and Jacobs, D. 2023. LD-ZNet: A Latent Diffusion Approach for Text-Based Image Segmentation. In *ICCV*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Sharma, G.; Liu, D.; Kalogerakis, E.; Maji, S.; Chaudhuri, S.; and Měch, R. 2020. ParSeNet: A Parametric Surface Fitting Network for 3D Point Clouds. In *ECCV*.

Tan, H.; Wu, S.; and Pi, J. 2023. Semantic Diffusion Network for Semantic Segmentation. In *NeurIPS*.

Tang, R.; Liu, L.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Stenetorp, P.; Lin, J.; and Ture, F. 2023. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In *ACL*.

Teichmann, M.; Weber, M.; Zöllner, J. M.; Cipolla, R.; and Urtasun, R. 2016. MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving. In *IV*.

Trémeau, A.; and Borel, N. 1997. A region growing and merging algorithm to color segmentation. *Pattern Recognition Letters.*

Voynov, A.; Morozov, S.; and Babenko, A. 2020. Object Segmentation Without Labels with Large-Scale Generative Models. In *ICML*.

Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; and Cattin, P. C. 2021. Diffusion Models for Implicit Image Segmentation Ensembles. In *MIDL*.

Wu, W.; Zhao, Y.; Chen, H.; Gu, Y.; Zhao, R.; He, Y.; Zhou, H.; Shou, M. Z.; and Shen, C. 2023a. DatasetDM: Synthesizing Data with Perception Annotations Using Diffusion Models. In *NeurIPS*.

Wu, W.; Zhao, Y.; Zhou, H.; Shou, M. Z.; and Shen, C. 2023b. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*.

Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; and Sun, J. 2018. Unified Perceptual Parsing for Scene Understanding. In *ECCV*.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *NeurIPS*.

Yang, J.; Gao, M.; Li, Z.; Gao, S.; Wang, F.; and Zheng, F. 2023. Track Anything: Segment Anything Meets Videos. arXiv:2304.11968.

Yu, F.; and Koltun, V. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*.

Zeiler, M. D.; and Fergus, R. 2013. Visualizing and Understanding Convolutional Networks. In *ECCV*.

Zhang, J.; Zheng, J.; and Cai, J. 2010. A diffusion approach to seeded image segmentation. In *CVPR*.

Zhang, Y.; Ling, H.; Gao, J.; Yin, K.; Lafleche, J.-F.; Barriuso, A.; Torralba, A.; and Fidler, S. 2021. DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort. In *CVPR*.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; and Zhang, L. 2021. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. In *CVPR*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing Through ADE20K Dataset. In *CVPR*.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment Everything Everywhere All at Once. In *NeurIPS*.