EMERDIFF: EMERGING PIXEL-LEVEL SEMANTIC KNOWLEDGE IN DIFFUSION MODELS

Koichi Namekata[1,2], **Amirmojtaba Sabour**[1,2], **Sanja Fidler**[1,2,3], **Seung Wook Kim**[1,2,3]
[1]University of Toronto, [2]Vector Institute, [3]NVIDIA
koichi.namekata@mail.utoronto.ca, {amsabour, fidler, seung}@cs.toronto.edu

ABSTRACT

Diffusion models have recently received increasing research attention for their remarkable transfer abilities in semantic segmentation tasks. However, generating fine-grained segmentation masks with diffusion models often requires additional training on annotated datasets, leaving it unclear to what extent pre-trained diffusion models alone understand the semantic relations of their generated images. To address this question, we leverage the semantic knowledge extracted from Stable Diffusion (SD) and aim to develop an image segmentor capable of generating fine-grained segmentation maps without any additional training. The primary difficulty stems from the fact that semantically meaningful feature maps typically exist only in the spatially lower-dimensional layers, which poses a challenge in directly extracting pixel-level semantic relations from these feature maps. To overcome this issue, our framework identifies semantic correspondences between image pixels and spatial locations of low-dimensional feature maps by exploiting SD's generation process and utilizes them for constructing image-resolution segmentation maps. In extensive experiments, the produced segmentation maps are demonstrated to be well delineated and capture detailed parts of the images, indicating the existence of highly accurate pixel-level semantic knowledge in diffusion models. *Project page:* https://kmcode1.github.io/Projects/EmerDiff/

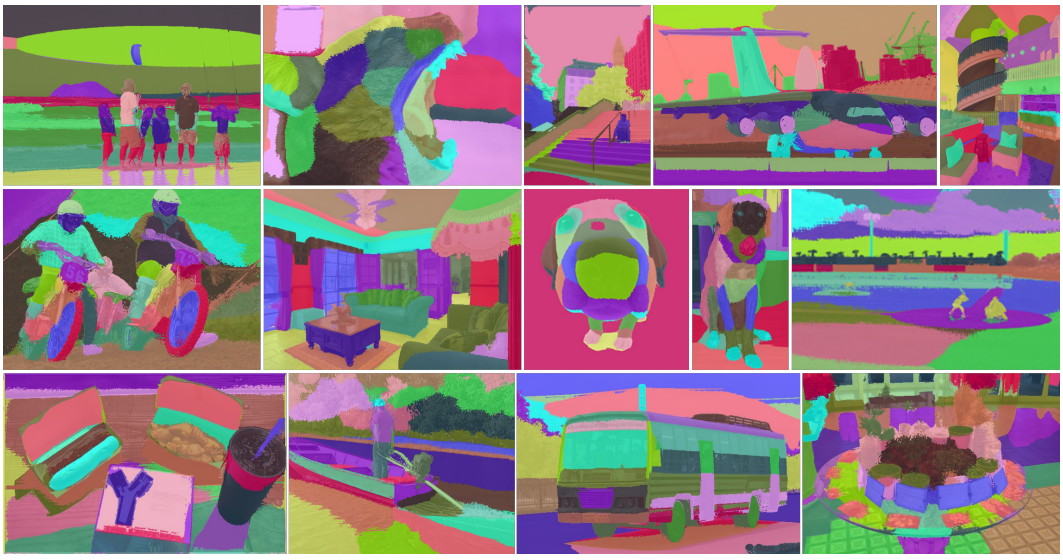

Figure 1: EmerDiff is an unsupervised image segmentor solely built on the semantic knowledge extracted from a pre-trained diffusion model. The obtained fine-detailed segmentation maps suggest the presence of highly accurate pixel-level semantic knowledge in diffusion models.

1 INTRODUCTION

In recent years, diffusion models (Ho et al., 2020; Dhariwal & Nichol, 2021) have emerged as the state-of-the-art generative models for synthesizing high-quality images. Notably, the internal representations of pre-trained diffusion models have been found semantically enriched, demonstrating impressive transfer abilities in semantic segmentation tasks (Baranchuk et al., 2022; Karazija et al., 2023; Li et al., 2023a; Xu et al., 2023a). However, the previous success in producing ne-grained segmentation maps often relies on incorporation of additional knowledge such as mask annotations (Baranchuk et al., 2022; Xu et al., 2023a) and hand-crafted priors (Karazija et al., 2023; Ma et al., 2023; Wu et al., 2023), leaving it unclear to what extent the pre-trained diffusion models alone understand the semantic relations of their generated images. To address this question, we present an unsupervised image segmentor that can generate ne-grained segmentation maps by solely leveraging the semantic knowledge extracted from a pre-trained diffusion model. Speci cally, our work is built on Stable Diffusion (Rombach et al., 2022), a large-scale text conditioned diffusion model capable of generating diverse high-resolution images.

The common unsupervised approach to capture semantic relations from a diffusion model is to apply k-means on its semantically meaningful feature maps. This technique has been demonstrated to produce cluster maps that are semantically aligned (Baranchuk et al., 2022; Patashnik et al., 2023; Xu et al., 2023a). While this method intuitively visualizes the semantic awareness of diffusion models, the obtained semantic relations are often coarse because of the fact that the semantically meaningful feature maps typically reside only in the spatially low-dimensional layers (Collins et al., 2020; Baranchuk et al., 2022; Luo et al., 2023). Nevertheless, diffusion models possess the ability to construct high-resolution images based on their semantic knowledge embedded in low-resolution feature maps. This motivates us to analyze how the semantically meaningful low-resolution feature maps in uence the output images through the generation process, which we hypothesize to be a crucial step for extracting pixel-level semantic knowledge from the diffusion models.

To this end, we investigate the in uence of local change in the values of the low-resolution feature maps on the pixel values of the generated images. Our key discovery is that when we perturb the values of a sub-region of low-resolution feature maps, the generated images are altered in a way that only the pixels semantically related to that sub-region are notably changed. Consequently, we can automatically identify the semantic correspondences between image pixels and a sub-region of low-dimensional feature maps by simply measuring the change in the pixel values.

Building on this insight, our proposed image segmentor can generate ne-grained segmentation maps without the need for any additional knowledge. First, we generate low-resolution segmentation maps (e.g., 16×16) by applying k-means on low-dimensional feature maps. Then, we build high-resolution segmentation maps (e.g., 512×512) in a top-down manner by mapping each image pixel to the most semantically corresponding low-resolution mask. These semantic correspondences are extracted from the diffusion models leveraging the aforementioned nding.

The effectiveness of our framework is extensively evaluated on multiple scene-centric datasets such as COCO-Stuff (Caesar et al., 2018), PASCAL-Context (Mottaghi et al., 2014), ADE20K (Zhou et al., 2019) and Cityscapes (Cordts et al., 2016) both qualitatively and quantitatively. Although the underlying diffusion model is never trained on annotated datasets, our framework produces segmentation maps that align surprisingly well with the detailed parts of the images, indicating the existence of highly accurate pixel-level semantic knowledge in the diffusion models.

2 RELATED WORKS

Generative models for semantic segmentation The use of generative models for semantic segmentation dates back to the era of GANs, where Collins et al. (2020) discovered applying k-means on the StyleGAN's (Karras et al., 2019) intermediate feature maps yields clusters aligning well with semantic objects. Following that, the prior works (Li et al., 2021; Tritrong et al., 2021; Xu & Zheng, 2021; Zhang et al., 2021; Li et al., 2022b) exploited such semantically meaningful feature maps to learn semantic segmentation with minimal supervision. Notably, diffusion models exhibit similar properties, where clusters of their intermediate feature maps consistently group semantic objects (Baranchuk et al., 2022; Xu et al., 2023a). These feature maps are utilized in various downstream tasks, including keypoint matching (Hedlin et al., 2023; Luo et al., 2023; Tang et al.,

2023a; Zhang et al., 2023) and semantic segmentation (Baranchuk et al., 2022; Li et al., 2023a; Xu et al., 2023a), outperforming GAN-based counterparts. Additionally, cross-attention layers of text-conditioned diffusion models (Balaji et al., 2022; Rombach et al., 2022; Saharia et al., 2022) are used for determining object layouts (Hertz et al., 2022; Patashnik et al., 2023; Tang et al., 2023b), where delimited layouts serve as (pseudo-) segmentation masks (Karazija et al., 2023; Ma et al., 2023; Wu et al., 2023). However, such semantically meaningful feature maps usually exist in low-dimensional layers, whose spatial resolution is significantly lower than the image resolution. To obtain upsampled and refined segmentation maps, prior literature has incorporated post-processing tools such as boundary refinement techniques (Kembhavi et al., 2011; Barron & Poole, 2016; Araslanov & Roth, 2020; Wang et al., 2023) which, however, relies on hand-crafted priors. In contrast, our framework successfully produces fine-grained segmentation masks (e.g. 512×512) from low-resolution feature maps (e.g. 16×16) without the need for any additional knowledge.

Unsupervised semantic segmentation. Unsupervised semantic segmentation is the task of grouping pixels of unlabeled images into semantically meaningful concepts without seeing annotated datasets. Previous studies have focused on either segmenting only salient objects (Van Gansbeke et al., 2021; Melas-Kyriazi et al., 2022a; Shin et al., 2022; Wang et al., 2022a; Simeoni et al., 2023; Zadaianchuk et al., 2023) or segmenting entire scenes (Ji et al., 2019; Cho et al., 2021; Hamilton et al., 2022; Seitzer et al., 2022; Wen et al., 2022; Yin et al., 2022; Li et al., 2023b), where our study is of the latter category. Typically, those frameworks are composed of two parts: an image encoder that produces pixel embeddings through self-supervised learning, and concept embeddings that are used to group pixels into a pre-defined number of semantic concepts. During inference, the pixels are classified into one of these concepts by matching their pixel embeddings with the closest concept embeddings. The current state-of-the-art approaches for scene segmentation (Hamilton et al., 2022; Seitzer et al., 2022; Wen et al., 2022) are built on the pre-trained self-supervised ViTs like DINO (Caron et al., 2021). There also exist several studies that utilize GANs' latent spaces for segmenting foreground objects (Voynov et al., 2020; Abdal et al., 2021; Melas-Kyriazi et al., 2022b; Feng et al., 2023; Oldfield et al., 2023); however, these approaches only work in narrow visual domains and are not applicable to scene-centric images.

Open-vocabulary semantic segmentation. Open-vocabulary semantic segmentation aims to segment images according to arbitrary user-defined vocabularies during inference. Models are typically trained with only text-image pairs (Xu et al., 2022a; Zhou et al., 2022; Cha et al., 2023; Mukhoti et al., 2023; Ranasinghe et al., 2023; Xu et al., 2023b), or combination of unlabeled/labeled annotations and text-image supervision (Ghiasi et al., 2022; Li et al., 2022a; Xu et al., 2022b; Liang et al., 2023; Xu et al., 2023a;c). The majority of these models are built on image encoders of pre-trained vision language models like CLIP (Radford et al., 2021), but learn to produce feature representations that exhibit better pixel-level alignment with text embeddings. However, the existing annotation-free models, which are trained without annotated datasets, tend to produce noisy segmentation masks. To overcome this issue, we integrate our framework into these models. Concretely, for each segmentation mask produced from our framework, we compute its mask embedding through their image encoders and classify them by the text embeddings. By combining our framework's segmentation abilities with their classification abilities, we achieve significantly better mIoU.

3 METHODS

As illustrated in Figure 2, our goal is to generate fine-grained segmentation maps by solely leveraging the semantic knowledge extracted from pre-trained diffusion models. To achieve this, we begin by generating low-resolution segmentation maps by applying k-means on the semantically meaningful low-dimensional feature maps (Section 3.2). Next, we construct image-resolution segmentation maps by mapping each image pixel to the most semantically corresponding low-resolution mask. To find the semantic correspondences between the image pixels and the masks, we exploit the diffusion model's mechanism of generating high-resolution images from their low-resolution feature maps (Section 3.3). In the following sections, we first provide an overview of the properties of diffusion models (Section 3.1) and then delve into further details of our approach.

3.1 PRELIMINARIES

Diffusion models are trained to generate images by taking successive denoising steps from pure Gaussian noise, where each denoising step is commonly performed with U-Net backbones (Ron-