Exploring Limits of Diffusion-Synthetic Training with Weakly Supervised Semantic Segmentation

Ryota Yoshihashi, Yuya Otsuka, Kenji Doi, Tomohiro Tanaka, and Hirokatsu Kataoka

LY Corporation

Abstract. The advance of generative models for images has inspired various training techniques for image recognition utilizing synthetic images. In semantic segmentation, one promising approach is extracting pseudo-masks from attention maps in text-to-image diffusion models, which enables real-image-and-annotation-free training. However, the pioneering training methods using the diffusion-synthetic images and pseudo-masks, e.g., DiffuMask have limitations in terms of mask quality, scalability, and ranges of applicable domains. To address these limitations, we propose a new framework to view diffusion-synthetic semantic segmentation training as a weakly supervised learning problem, where we utilize potentially inaccurate attentive information within the generative model as supervision. Motivated by this perspective, we first introduce reliability-aware robust training, originally used as a classifier-based WSSS method, with modification to handle generative attentions. Additionally, we propose techniques to boost the weakly supervised synthetic training: we introduce prompt augmentation by synonym-and-hyponym replacement, which is data augmentation to the prompt text set to scale up and diversify training images with limited text resources. Finally, LoRA-based adaptation of Stable Diffusion enables the transfer to a distant domain, e.g., auto-driving images. Experiments in PASCAL VOC, ImageNet-S, and Cityscapes show that our method effectively closes gap between real and synthetic training in semantic segmentation. Our code will be available at https://github.com/yahoojapan/attn2mask.

Keywords: Diffusion model · Semantic segmentation · Weakly supervised learning · Diffusion-synthetic training

1 Introduction

Advanced generative models have enabled high-fidelity and controllable synthesis of images. Diffusion models [25] can generate high-quality images without adversarial training [17], which often makes the optimization tricky. Large-scale image-caption pair datasets [62] enabled text-to-image (text2img) generation [49,54,57], which is useful to control generated image contents by using free-form textual prompts. These generative trends inspired research in discriminative learning to train models only using generated images without relying on manually collected and annotated image datasets [14,19,61,68,69]. This opportunity to exploit *generative models as data* [30,80] requires rethinking existing training strategies to better fit the synthetic training settings.

In semantic segmentation, it is essential to generate pixel-wise semantic labels in addition to images for enabling synthetic training. Fortunately, text2img diffusion models naturally can extract pixel-semantic labels by utilizing their internal text-image cross-attention layers [76]. In the generation processes, the cross-attention layers are used to relate the text tokens to the elements of latent image representations to reflect the prompt. In other words, the cross-attention layers solve the grounding between words and pixels in generated images [66], and they can serve as sources of localization information [22,67,87]. These text-image cross-attention maps, when paired with generated images, can be used as supervisory signals for segmentation. A pioneering work in this direction is DiffuMask [76], which generates training images and labels for each class of the target benchmark (e.g., PASCAL VOC [13]) with Stable Diffusion (SD) [57].

However, such a synthetically trained segmentation model currently lags behind the real-image-trained counterpart in accuracy. One reason for this is the generated image quality, but it may be a matter of time until this problem is solved by the drastic and continuing improvements of generative models. Another possible reason is the sensitivity of segmentation training to low-quality labels generated with attention, which should be cared for by the segmentation side. More specifically, we analyze the previous method DiffuMask [76]'s limitations in the following three aspects: First, the quality of synthetic labels, which is not always as accurate as human-annotated ones, affects the segmentation models' performances directly. Second, the image synthesis largely relies on template-based artificial prompts only related to one among the target classes (e.g., 20 classes in PASCAL VOC), which limits the scale and diversity of training data. Third, applying the segmentation models to domains distant from the original SD's training domain, i.e., web-crawled images, is hard. DiffuMask was only applicable to small subclasses of all defined classes when it was transferred to Cityscapes [10], a dataset for urban auto-driving.

We hypothesize that the first and largest part of these limitations comes from fully supervised segmentation training [5, 8, 45] using the synthetic data that are not yet perfect. In contrast, we frame the synthetic training as a weakly supervised task where we need to train accurate segmentation from possibly inaccurate generated labels. This framing is reasonable because text2img generators, including SD, can be seen as text-supervised models, and synthetic training using them can be regarded as text-supervised weakly supervised learning for segmentation. Weakly supervised semantic segmentation (WSSS) with real images has been well studied [51, 58, 86], and especially, we incorporate the idea of robust co-training for WSSS [58] with modifications for synthetic training. One notable difference is that conventional WSSS mainly utilized the classifiers' activation maps (CAMs), but we replace them with generative attentions, whose statistics are different as discussed later. However, we successfully overcome the difference by tailoring an adaptive threshold strategy for the SD's generative attentions.

We further explore the limitations from the two perspectives: scalability and range of application. To easily scale up the size of the synthetic training data, we propose prompt augmentation by synonym-and-hyponym replacement, which is

a data augmentation technique in the prompt space rather than in the image space and is useful to inflate the limited text corpus to seed more generated images. Such augmentations are often seen in language-based tasks [4, 73], but we discover its applicability as image augmentation in text-to-image-based synthetic training for enhancing vision models. To widen the range of applications, we further incorporate an adaptation technique with LoRA [27], which quickly finetunes SD with a small amount of target-domain real unannotated images. Here, our finding is that SD does not lose its mask-producing ability after LoRA-based finetuning even though the target-domain images are not paired with textual prompts, which successfully perform unsupervised domain adaptation.

Incorporating these ideas, we developed our diffusion-synthetically trained semantic segmentation method *Attn2mask*. In experiments, we verify Attn2mask's effectiveness by achieving 62.2 mIoU in the PASCAL VOC segmentation task, which is accurate for not using any real images or human annotation. Attn2mask is also applicable to a larger-scale scenario in ImageNet-S 1k-class segmentation, where it performs sufficiently close to the semi-supervised BigDatasetGAN [38] generator. Furthermore, the adaptation of SD is shown to be beneficial for distant-domain transfer for the Cityscapes driving images, although it slightly differs from the spirit of real-image-free training.

Our contributions are summarized as follows:

Conceptual contribution. We connect the ideas of diffusion-synthetic training and WSSS for the first time. Concretely, we demonstrate that even attentive masks from the strong SD model can be regarded as *weak* supervision and have room for improvement with WSSS training techniques.

Technical contribution. We propose a bag of approaches to push the limits of diffusion-synthetic WSSS, namely, adaptive threshold for generative attention maps, prompt augmentation to boost the diversity of synthetic datasets, and domain-adaptive LoRA to make SD applicable to distant-domain tasks. The effects of each of them are validated by the experiments.

Experimental contribution. We demonstrate Attn2mask's effectiveness equipped with these ideas in the experiments, especially by achieving 62.2 mIoU with ResNet50-based networks, which outperforms the existing state of the arts in synthetically trained semantic segmentation without human-annotated labels.

2 Related work

Segmentation with diffusion models Studies have tried to leverage the diffusion models' excellent image-generation performance in segmentation. There are two lines of research: reusing diffusion U-Net for segmentation tasks and generating training image-label pairs for segmentation. For the former line, denoising diffusion probabilistic models (DDPMs) can be used as pre-trained feature extractors for semantic segmentation [2]. ODISE [77] and VPD [84] similarly finetune a text2img diffusion model for segmentation, but it enables phrase-based segmentation by reusing a text encoder in the text2img model. UniGS [52] finetunes the diffusion model for multi-task generation and segmentation. Similar approaches are seen in single-object localization [85] or object discovery [46]. These usages of text2img models as visual-linguistic backbones can be seen as

a generative counterpart of contrastive visual-linguistic foundation models, e.g., CLIP [53], which also can be repurposed for segmentation tasks [55]. However, all of the aforementioned methods rely on manually annotated pixel labels for training. Very recently, diffusion-based training-free methods were proposed [33, 70]. While they are attractive by omitting the additional training phase, their segmentation accuracies tend to be suboptimal; in contrast, we emphasize that robust re-training based on WSSS techniques is our source of improvement.

We call the latter line of research *diffusion-synthetic training*. For example, Grounded Diffusion [41] trains image-and-mask generation models by leveraging a pre-trained Mask R-CNN with pixel labels. FreeMask [78] augmented real image-and-mask datasets using mask-to-image generation. We refer to these kinds of approaches *semi-supervised* mask generation because of their dependency on annotated training images on at least partial classes. DiffuMask [76] is closer to ours in the sense of *weakly supervised* mask generation, where masks are automatically acquired via text2img generation training, but it only supports single-object image generation. Dataset Diffusion [48] is another work in this line. It enables multi-class image-mask generation using real image captions. In the concept level, ours differs in that we frame the problem in WSSS to focus on improving possibly inaccurate generated labels in the segmentation training phase, while the two studies adopted fully supervised segmentation in the downstream training. In more detail, 1) we use prompt augmentation to acquire more diversified and contextualized prompts instead of template-based or real-caption-based prompts, and 2) we omit iterative self-training [48] or cross-validation-based synthesized image curation [76] thanks to the co-training that is robust even with single-step training.

Training with generated images The idea of exploiting free annotations with synthesized images dates back to before the large-scale generative models: for example, computer-graphics-based dataset creation using game engines was examined [56, 59]. Cut-paste-based learning [12, 16] can be seen as a simpler form of the dataset-generation methods. More primitively, even segmentation of mathematically drawn contours can help real segmentation as a pretraining task [34, 64]. However, such rendered images tended to have gaps from real ones, which needed supervised retraining or unsupervised adaptation techniques [26] to utilize them fully to achieve preferable real-environment performances. In contrast, we exploit modern generative models' high-fidelity generation ability that narrows the gap without any adaptation at least in generic settings such as PASCAL VOC. Prior to the diffusion models, GAN-based image-and-mask generation was also tried [38] but manual annotation was still required for partial classes to fine-tune the GANs for mask generation. In addition to segmentation, classification [61] and unsupervised representation learning [69, 82] with generated images have been studied, and they report performances gradually approaching those of real-dataset-based counterparts, which is motivating for studies in synthetic training including ours.

Weakly supervised semantic segmentation Mainstream WSSS approaches exploit image-level class labels. Two-stage training that first trains an image-level classifier and then re-trains another segmentation model using the classifier's

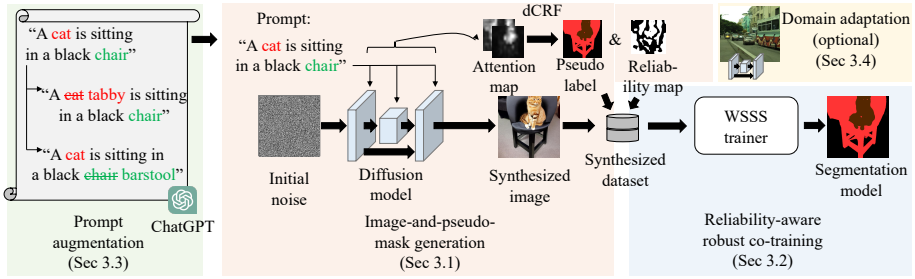

Fig. 1: Overview of the proposed method *Attn2mask*. The major parts are image-and-pseudo-mask generation and reliability-aware robust co-training. Additionally, we adopt prompt augmentation, a data-augmentation technique for prompts. Domain adaptation is optionally supported for far-domain transfer.

class-activation maps (CAM) [86] as weak labels is the most straightforward but still well-used strategy. Among the two-stage methods, existing studies attempted techniques for better CAM generation [7,31,40,72] and robust segmentation training against weak labels [39,43,58]. From the viewpoint of WSSS, we replace classifier-based CAMs with generative attentions combined with generated images. In addition, we incorporate the latest robust segmentation-training method [58], which we find to be beneficial in our training scenario. A preprint concurrent with ours leveraged ControlNet [83] for data augmentation in WSSS [75], but this approach is largely different from ours in terms of not supporting pure synthetic training without real images. Recent methods integrated the Segment-Anything Model (SAM) [35] in WSSS [65, 79]. SAM in itself is a strongly supervised for segmentation because its pretraining involves human pixel-annotation. While ours similarly relies on SD, it is weakly supervised for segmentation since it depends on image-text pairs but no pixel-level intervention; one can be less reluctant to call our method weakly supervised even considering the pre-training.

Unsupervised segmentation and text-based segmentation Recently, segmentation studies diverged in various training strategies. For example, unsupervised semantic segmentation does not use any form of supervision in the downstream segmentation training [9,81]. Our Attn2mask may be seen as unsupervised in the sense that it is annotation-free in the downstream domains, but it operates under the stricter constraint of unavailability of downstream-domain (real) images than usual unsupervised methods. Text-based segmentation exploits vision-language pretraining [55, 63, 87], which often transfers well in the downstream segmentation tasks without fine-tuning. Attn2mask is related to these methods in the usage of vision-language pretraining except for the difference between generative (i.e., SD) and discriminative (i.e., CLIP) pretraining.

3 Proposed method Attn2mask

An overview of Attn2mask is shown in Fig.1. Its major parts are image-and-pseudo-mask generation and segmentation training. In the segmentation-training, we utilize a robust co-training strategy to handle possibly inaccurate generated pseudo-masks. Additionally, we adopt prompt augmentation, a data-augmentation technique for prompts to generate images rather than to the generated images.

3.1 Image-and-pseudo-mask generation

To extract semantic regions from diffusion-synthesized images, we need to compute the contributions of each word in the prompts on each location in images. Fortunately, image-text cross-attention, commonly used in most diffusion models including SD, is naturally available for this purpose. Transformer-style attention modules can be written as follows:

$$\text{Transformer}(Q, K, V) = A(Q, K)V, \tag{1}$$

$$A(Q, K) \quad = \text{Softmax}(\frac{QK^T}{\sqrt{\dim(K)}}), \tag{2}$$

where Q, K, and V are the query, key, and value, each of which is a sequence of vectors. In using this as text-to-image cross-attention, Q is computed from the image embedding $\boldsymbol{f} \in \mathbb{R}^{W \times H \times C}$ whose size is $W \times H$ and dimensionality is C, and K and V are computed from the text embedding $\boldsymbol{s} \in \mathbb{R}^{L \times D}$ whose length is L and dimensionality is D. This makes the attention maps' dimensionality

$$A(Q(\boldsymbol{s}), K(\boldsymbol{f})) \in \mathbb{R}^{W \times H \times L}, \tag{3}$$

where the element at (x, y, t) can be interpreted as the amount of contribution from the t-th text token to the location in image coordinate (x, y).

From the attention maps A, we extract channels that are relevant to the classes of interest to use them as supervisory masks. Given a token sequence of the input prompt, we denote the positions of the relevant tokens of class c by $\tau_c = [t_{c1}, t_{c2}, ..., t_{cn_c}]$, allowing multiple occurrences of the relevant tokens for a single class. The "relevance" of words and classes can be set manually by any rule-based matching, and in our implementation, we list singular and plural forms of synonyms for a class's display name and extracted exact matches. This relevance-based attention extraction can be denoted as

$$A_c = \frac{1}{n_c} \sum_{t \in \tau_c} A(Q(\boldsymbol{s}), K(\boldsymbol{f}))_t \in \mathbb{R}^{W \times H}, \tag{4}$$

where we denote the operation to extract t-th channels of the attention map by $A(\cdot, \cdot)_t$.

We add one more probability map for the background as a class defined by

$$A_0(x, y) = 1.0 - \max_{c \in [1, N]} A_c(x, y) - \beta, \tag{5}$$

which represents the absence of activations for any other classes. Here, N is the number of (non-background) classes, and β is a hyperparameter to control the strength of the background prior; larger β brings smaller portions of the background area in labeling results. The motivation behind this special treatment for the backgrounds is that they tend to be generated implicitly without concrete indication in the prompts, which makes class-word-based aggregation in Eq. 4 difficult.

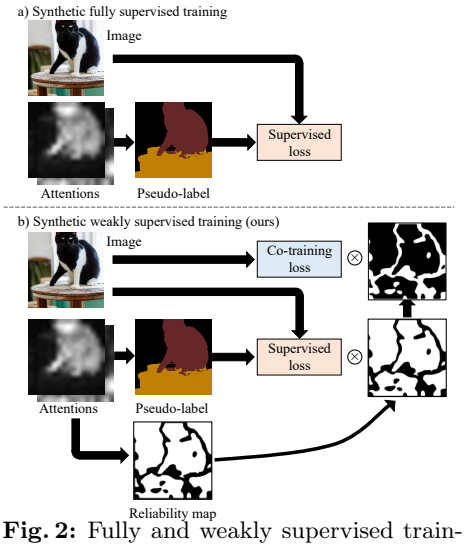

Fig. 2: Fully and weakly supervised training schemes with synthetic pseudo labels.

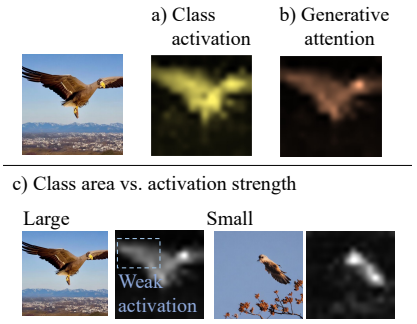

Fig. 3: Comparison between a) CAM and b) generative attentions. The generative attentions is prone to cause weak activations simply by the large area that a class (in the examples *bird*) occupies. We mitigate this problem with adaptive thresholding.

After attention maps $A = [A_0, A_1, ..., A_N]$ are gathered, we apply binarization to them to acquire multi-class discrete pseudo-labels. We use the densely connected conditional random field (dCRF) [36], a non-learning-based labeling algorithm. This is also useful to improve mask quality by considering the color similarity and local smoothness in labeling. Although the prior work [76] applied dCRF after threshold-based binarization of FG/BG masks, we use dCRF in multi-class, continuous-probability-based labeling that is the original usage in [36]. These generation processes are iterated for a given set of prompts, and generated images and masks are stored in the disk as a dataset.

3.2 Reliability-aware robust training

Once our images and pseudo-labels are generated, any supervised segmentation models are generally trainable with them. However, naive training of supervised learners may suffer since the attention-based pseudo-labels are not always perfectly accurate; robust learners that are tolerant of label noises would help. Thus, we adopt a robust training algorithm from BECO [58], a recent WSSS method with modification to utilize generative attention maps as shown in Fig. 2.

BECO is an adaptive co-training method exploiting reliability maps of the pseudo-labels: it performs supervised learning with full trust in the pseudo-labels within regions where the pseudo-labels are confident and performs co-training-based consistency regularization in regions where the pseudo-labels are unreliable. In the original work, BECO was trained with reliability maps generated by processing the classifier's confidence score maps. Rather than additionally train a classifier, a naive approach might be binarizing the attention maps with a constant threshold to use as the reliability map as follows:

$$R(x,y) = \begin{cases} 1 \text{ if } \max_{c\in[0,N]} A_c(x,y) \geq r \\ 0 \text{ otherwise,} \end{cases} \tag{6}$$

where r is a hyperparameter for the reliability threshold.

However, there is a difference in sources of pseudo-masks that poses a challenge to application of the reliability-based co-training: conventional WSSS has utilized CAM computed from a classifier, which is often specially trained to produce the activation maps that are suitable for segmentation [7, 40]. The CAM is discriminatively optimized and therefore it is intuitive that we can interpret their raw values as reliability. In contrast, we found that generative attentions in SD do not necessarily have strong activations discriminative regions. A conspicuous tendency is that a larger object is harder to cover with high activation values simply because the values are distributed across a broader region as seen in in Fig. 3 c, which may be harmful by discarding a large portion of properly generated and pseudo-masked regions with a naive thresholding.

Thus, we newly developed an adaptive threshold strategy that considers the attention strength, spatial distributions of the attention, and smoothed label assignments by dCRF in a joint manner as follows:

$$S = \mathrm{dCRF}(A), \tag{7}$$

$$r(A, c) = \alpha \frac{\sum_{(x,y)\in\{(x,y)|S(x,y)=c\}} A_c(x, y)}{\sum_{(x,y)\in\{(x,y)|S(x,y)=c\}} 1}, \tag{8}$$

where $S \in \{0, 1, 2, ..., N\}^{W\times H}$ denotes labels assigned by dCRF, $r(A, c) \in [0, 1]^{W\times H}$ denotes the reliability-threshold maps, and $c \in 0, 1, 2, ..., N$ denotes the class of interest. A hyperparameter α is introduced to control the overall hardness of the thresholds. This threshold becomes smaller when the area of the class of interest used in the denominator $\sum_{(x,y)\in\{(x,y)|S(x,y)=c\}} 1$ is larger, and is suitable to mitigate discarding the weak attention activations when the class areas are large. These adaptive threshold maps are applied to the attention pseudo-masks with w.r.t. dCRF's selection of class labels in each location as

$$R(x, y) = \begin{cases} 1 & \text{if } A_{S(x,y)}(x, y) \geq r(A, S(x, y)) \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

and in this way we acquired the reliability map to switch on/off the co-training.

3.3 Prompt augmentation

The abovementioned processes rely on a set of prompts to generate many diversified training samples. However, preparing a number of proper texts may be additional labor even though it is less hard than pixel annotation. Existing work utilized template-based prompt generation [76] or real captions from a captioning dataset [48], but they limited the total numbers of available training images to less than 100k.

We take a simple but effective strategy for gathering prompt texts; we first curate real captions related to the target classes and next automatically augment them with *synonym-and-hyponym replacement*. Synonym replacement [74] is a technique for text data augmentation to replace words randomly with other words that have the same or similar meanings, for example "bicycle" by "bike".

We additionally use the hyponyms s, for example "owl" for "bird", for replacement because they are useful to diversify generated images. Lists of synonyms and hyponyms for the 20 VOC classes are easily collected by asking ChatGPT [50] "`please raise fifteen examples of synonyms or hyponyms for ${class_names}`". We utilize the prompt augmentation in combination with image-space data augmentation.

The advantage of our prompt augmentation strategy is that it scales up a text corpus in a combinatorial manner: Given T texts and C synonyms for a class, our prompt augmentation can generate TC variations of prompts. Reviewing existing prompt strategies, Dataset Diffusion [48] relied on the raw texts with the size T, and DiffuMask [76] added simple templates like "a photo of [synonym]," which makes the corpus scale $T + C$; both were bottlenecked with the corpus sizes.

3.4 Domain adaptive image-mask generation

While image-mask generation with off-the-shelf SD generally performs satisfactorily, in out-of-domain applications, it may suffer from the domain gap between the generative model and the downstream tasks. We notice that this is the case, for example, in applying Attn2mask to driving images in Cityscapes [10].

To make Attn2mask applicable to new domains, we develop an adaptation method using unlabeled training images. Here, we repurpose the instantiated generation technique DreamBooth [60] for domain adaptation. For a brief review, DreamBooth learns an instance token (often denoted by [V]) from a relatively small set of instance images. After fine-tuning with the instance-image set, [V] is usable as a new word in prompts to indicate the instance. Using DreamBooth, we instead learn a domain token; by training [V] not for a specific instance but for general training images in a new domain, SD can adapt to the domain, while maintaining the ability to reflect other prompt words to the image contents.

In implementation, we especially use DreamBooth-LoRA [28], which trains a low-rank adapter [27] that is a small sub-network instead of fine-tuning the whole network. This enables fast and stable training with a relatively small target-domain training set.

4 Experiments

We evaluate semantic-segmentation models trained with Attn2mask on three publicly available datasets: PASCAL VOC 2012 (VOC12) [13], ImageNet-S [15], and Cityscapes [10]. We conduct the main experiments in VOC12, which has been the de facto standard in semantic-segmentation evaluation. While fully supervised learning in VOC seems close to saturation in terms of accuracy, it is challenging enough to tackle without full annotation and real images. Additionally, we evaluate Attn2mask's generalizability to 1) a larger-scale segmentation task with ImageNet-S, which was built on ImageNet-1k [11] and 2) a more task-specific scenario with Cityscapes, a dataset toward urban auto-driving.

4.1 Dataset generation

We used SD 1.4 [29] trained on LAION-2B [62] without modification except in the domain-adaptation settings. We generated 512 × 512-resolution images.

Table 1: Segmentation IoU on VOC12 with the ResNet50 backbone. Our Attn2mask outperformed DiffuMask in the mean.

Method	bg	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table
Real-supervised	94.4	90.0	42.4	82.1	70.5	75.8	93.4	88.1	90.7	36.5	86.5	67.2
DiffuMask [76]	n/a	**80.7**	n/a	**86.7**	56.9	n/a	81.2	74.2	79.3	14.7	63.4	n/a
Attn2mask (ours)	88.7	65.7	34.7	82.5	**64.7**	62.5	**87.0**	**76.0**	**83.2**	**25.0**	**65.3**	49.3

Method	dog	horse	motorbike	person	plant	sheep	sofa	train	tv	mean
Real-supervised	86.0	90.2	87.2	85.0	68.4	88.5	57.3	84.1	78.6	78.2
DiffuMask [76]	65.1	64.6	n/a	**71.0**	n/a	64.7	**27.8**	n/a	n/a	57.4
Attn2mask (ours)	**73.0**	**65.2**	73.6	64.8	35.2	**69.0**	13.6	62.4	58.3	**62.2**

The attention maps had smaller resolutions but were resized to match the image resolution. SD's U-Net had multiple cross-attention layers and they were averaged after resizing. We used DDPM [25] sampler with 100 timesteps, and attention maps were taken at the 50th step. Generation of a sample took around eight seconds with an NVIDIA V100 GPU. Post-processing was conducted using dCRF [21] with the default hyperparameters. The whole generation process was done twice with different random seeds to hold out the validation set.

The prompt set for VOC was created from the image descriptions from COCO captions [6]. We curated captions that included VOC class names and their major synonyms that we manually listed up, which made 88,561 captions in total. These captions were doubled by prompt augmentation, and finally 80k were selected by CLIPScore [23] filtering. The prompt set for ImageNet-S was created by a template "`a photo of ${class_name} with a background`". The variable `${class_name}` was randomly chosen from the multiple-defined synonyms per class. This made the number of possible varieties of the prompts significantly fewer than the needed samples per class (around 100 images/class) for our experiments, and thus we increased the number of samples by using multiple random seeds per prompt. The suffix "`with a background`" helps to reduce the generation of full-foreground images, which is less useful for training.

For Cityscapes, we used a variable-length template "`a photo of ${class_names[0]}, ${class_names[1]}, ... , ${class_names[n]} in cityscape by a dashboard camera`", where `class_names` is a random variable-length list of Cityscapes class names. A number of the class names in Cityscapes were not everyday words and thus may not be reflected in the generation, so we replaced "`terrain`" with "`sand`" and "`vegetation`" with "`plant`", and generated 12k samples. In the adaptation experiments, we used a modified template "`a photo of ${class_names[0]}, ${class_names[1]}, ... , ${class_names[n]} in sks cityscape by a dashboard camera`", where "`sks`" was the reserved word for the domain token [V] in DreamBooth fine-tuning. More details of this fine-tuning are seen in Appendix.

4.2 Configurations of our model

We implemented two versions of Attn2mask with the same SD but segmentation models with different backbones. We used DeepLabv3+ [5] with the ResNet50 [20] and Swin-B [44] backbone from MMSeg [47]. For training, we used the BECO

Table 2: Comparisons with the existing diffusion-synthetic semantic segmentation methods on the VOC12 val set. †: our reproduction.

Method	Backbone	mIoU
With ordinary-size backbones		
DiffuMask [76]	ResNet50	57.4
Dataset Diffusion [48]	ResNet50	61.6
Attn2mask (ours)	ResNet50	**62.2**
With larger backbones		
DiffuMask [76]	Swin-B	70.6
Dataset Diffusion [48]	ResNet101	64.8
Dataset Diffusion †	Swin-B	69.5
Attn2mask (ours)	Swin-B	**71.0**

Table 3: Synthesized dataset quality measured with FID and KID with VOC.

FID↓	Image	Mask
Ours	**28.2**	**77.0**
Dataset Diffusion	37.7	134.9

KID↓	Image	Mask
Ours	**0.021**	**0.060**
Dataset Diffusion	0.024	0.14

codebase [58]. The reliability threshold coefficient α was set to 1.0, and β was set to 0.1. Finally, we fine-tuned SD using DreamBooth-LoRA [28] for transfer to Cityscapes, which we refer to as Attn2mask-LoRA. We set the base learning rate to 0.01, weight decay to 1e-4, and total batch size to 32 on 2 GPUs.

4.3 Evaluation protocol

For PASCAL VOC and Cityscapes, we followed the official protocol for supervised learners, which is applicable without modification for WS learners, including ours. We used the mean intersection-over-union (mIoU) metric. For ImageNet-S, we derived a foreground/background (FG/BG) segmentation task in addition to the original 919-class segmentation by extracting masks that have the same class labels as their whole images. This was intended to replicate the evaluation protocol of BigDatasetGAN [38], in which the authors built their own ImageNet-based segmentation benchmark, but it remains unpublished.

4.4 Results

Main result on VOC (Tab. 1 and 2). Attn2mask marked mIoU of 62.2%, which is, of course, worse than the results by supervised training with the real images but reasonably well for not using real images or manual annotation at all. It outperformed DiffuMask [76], which is an important prior work in diffusion-synthetic semantic segmentation without real images with a margin of 4.8% in the mIoU. In particular, Attn2mask performs closely to the supervised counterpart in relatively easy classes such as *bird* (82.1% v.s. 82.5%), *bus* (93.4% v.s. 87.0%), and *cat* (90.7% v.s. 83.2%) or contrarily difficult classes where the supervised models do not work well, such as *bicycle* (42.4% v.s. 34.7%) or *chair* (36.5% v.s. 25.0%). However, we also acknowledge that there exists significant degradation of IoU in some classes such as *plant* and *sofa* compared with the real-image training. We observed that such classes tend to be generated as parts of backgrounds without being explicitly prompted. This made noise for training these classes, which is the major limitation of our synthetic training. Table 2 shows additional comparisons using different backbones. Attn2mask outperformed all other diffusion-synthetic semantic segmentation training methods without human annotation [48, 76], with ResNet50 and Swin-B settings. Table 3 shows quality criteria FID [24] and KID [3] of our images and masks w.r.t real VOC,

Table 4: Comparisons with real-image-based WSSS methods on VOC12 val.

Method	Training data Real	Synthetic	mIoU
Pseudo-mask [40]	✓		70.0
ReCAM [7]	✓		70.5
BECO [58]	✓		73.7
CLIP-ES [42]	✓		73.8
WSSS-SAM [65]	✓		77.2
MARS [32]	✓		**77.7**
Attn2mask (ours)	✓		71.7
Attn2mask (ours)		✓	71.0
Attn2mask (ours)	✓	✓	**75.6**

Table 5: Ablative analyses of Attn2mask. Evaluated with VOC12.

	mIoU
Baseline with ResNet50	44.6
+ dCRF	51.2
+ Co-training	58.4
+ Adaptive threshold	58.9
+ Prompt augmentation	59.8
+ CLIP filtering	61.3
+ TTA	62.2
+ Swin-B backbone	71.0
+ Real images	75.6

Table 6: Results in the ImageNet-S FG/BG segmentation evaluated with mIoU. Attn2mask enabled annotation-free large-scale segmentation.

Method	#Anno-tations	#Images 10k	100k
Real-Supervised	10k	82.1	–
BigDatasetGAN [38]	5k	74.1	74.4
Attn2mask (ours)	0	63.7	66.0

Table 7: Results in the ImageNet-S 919-class segmentation with the 100k-image settings.

Method	#Anno-tations	mIoU
Real-Supervised	10k	25.7
BigDatasetGAN [38]	5k	19.7
Attn2mask (ours)	0	13.3
PASS$_s$ [15]	0	11.5

which shows superior label quality of ours to Dataset Diffusion, while the image quality is similar depending on the same SD.

Comparisons with real-image-based WSSS (Tab. 4). The WSSS methods listed here used real VOCaug training images. Our Attn2mask performs closely to existing real-image-based WSSS before 2022 [7, 40] even without using the real images. While our primary interest is in pure synthetic training, one may be interested in whether the results have room for improvement with the help of real training images. Then, we added the real images from the VOC train set with pseudo-labels created by CAM and IRN [1] that was distributed along with BECO codebase, and this achieved mIoU 75.6%, which was competitive in the WSSS setting. It is slightly behind of MARS [32] and WSSS-SAM [65], but MARS used a complex object removal strategy and WSSS-SAM used pixel-supervised segment anything model (SAM) [35] for pseudo-label generation.

Ablative analyses (Tab. 5). While all of the added modules contributed to the segmentation performance, a particularly large gain was from the robust co-training. We also used test-time augmentation (TTA) and replacement by a larger backbone, which are common techniques in segmentation, for fair comparisons with prior work [48, 76].

Results on ImageNet-S (Tab. 6 and 7). Table 6 shows the results in ImageNet-S FG/BG segmentation. We compared Attn2mask with the same network trained with data from BigDatasetGAN [38]. BigDatasetGAN is a semi-supervised generator that relies on manual annotation of generated images to fine-tune GAN for mask generation. Attn2mask performed competitively without using pixel-label annotations at all, which is encouraging because shows the

ability to freely increase classes within the range Stable Diffusion can generate. In ImageNet-S full-class segmentation, a similar tendency was observed despite lower scores with all of the methods due to the hardness of the multi-class segmentation with the extreme numbers of classes, as shown in Table 7.

Domain-adaptative LoRA in Cityscapes (Tab. 8). LoRA-based adaptation of Attn2mask significantly outperformed the original version and full-finetuning version (Attn2mask-FT), which shows the popular SD's adaptation technique is also useful as an adaptive synthetic training technique. In contrast to segmenting object images such as VOC, holistic scene understanding in Cityscapes is extremely challenging for WSSS methods using image-level labels. Hence, unsupervised segmentation methods [9, 18, 63] are rather intensively developed and we placed reference mIoU values from them. Attn2mask performs similarly to the existing unsupervised methods even with the significant domain gap between web-crawled training images of SD and the driving images. We found the LoRA-based adaptation contributed to the segmentation accuracy, which is encouraging for future development of domain-adaptive diffusion-based segmentation methods. Attn2mask outperformed the strong unsupervised segmentation methods MaskCLIP [87], PiCIE [9] and, STEGO [18] in both settings with and without adaptation. ReCo [63] outperformed Attn2mask in the accuracy metric but relied on test-time heavy computation; it used retrieval of relevant images for co-segmentation.

Table 8: Results in Cityscapes with and without unsupervised adaptation.

Method	Adapt.	Acc.	mIoU
Attn2mask (ours)		50.6	21.1
Attn2mask-FT (ours)	✓	64.1	23.8
Attn2mask-LoRA (ours)	✓	75.8	**25.8**
MaskCLIP [87]		35.9	10.0
MDC [9]		–	7.0
PiCIE [9]		–	9.7
ReCo [63]		74.6	19.3
DiffSeg [67]		76.0	21.2
MDC [9]	✓	47.9	7.1
PiCIE [9]	✓	40.7	12.3
STEGO [18]	✓	73.2	21.0
ReCo+ [63]	✓	**83.7**	24.2
Deep clustering [71]	✓	−−	24.2

4.5 Visualizations

Figure 4 shows examples of generated VOC training images and labels by Attn2mask. In addition to the natural-looking object images, labels are highly accurate except for several missing detailed parts and over-masking of backgrounds where contextual connections exist, e.g., shadows of objects or waves around boats. We also noticed there exist severe failure cases such as masks not covering whole objects (bottom left) or mismatches of object classes and labels (bottom right). More examples will be seen in Supplementary Material.

Figure 5 visualizes synthesized images and attention maps before and after DreamBooth-LoRA-based adaptation. Before adaptation, image contents and styles were not well-aligned with Cityscapes and the naturalness of images was also somewhat harmed due to generation in a distant domain from the LAION dataset. After adaptation, generated images look enough like the

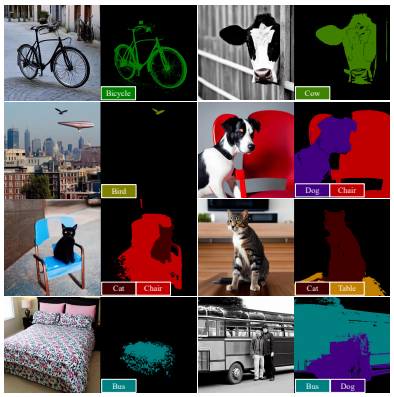

Fig. 4: Examples of our training data generated for VOC.

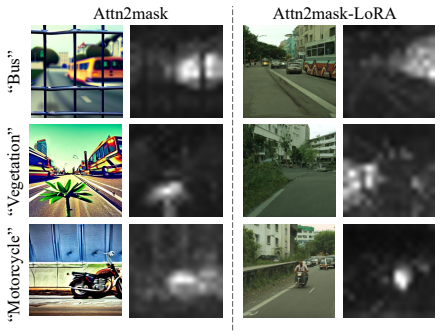

Fig. 5: Examples of images and attention maps before and after LoRA-based adaptations in Cityscapes.

images in Cityscapes, while attention maps were kept focused on the objects of interest. Interestingly, the correspondence between attention maps and objects is not corrupted after fine-tuning with the Cityscapes images that are not paired with any captions, which might be because DreamBooth-LoRA reuses the prior knowledge acquired in the original SD well.

5 Conclusion and discussion

We presented a real-image-and-annotation-free semantic-segmentation method Attn2mask. Although synthetic training did not outperform real-image-based counterparts to the extent we studied, Attn2mask worked surprisingly well for its purely real-image-free segmentation training, and the gaps between the real and synthetic performances were non-negligibly narrow by incorporating ideas from WSSS literature. To summarize our findings, 1) generative attentions are not always accurate, but we have chances to correct them in the segmentation-training phase by robust co-training (Sec. 3.2, Tab. 1 and 2). 2) Stable Diffusion's ability to synthesize diverse samples can be exploited for segmentation via prompt augmentation or ImageNet-class large-scale generation (Sec. 3.3). 3) In LoRA-based adaptation, Stable Diffusion does not lose its mask-generation ability and thus it is useful for transferring the knowledge to far domains (Sec. 3.4 and Tab. 8). The rising SD-synthetic training may cast a question like "Is SD all we need for WSSS?" to the community. We hope that our finding "WSSS and SD work better together" lays the groundwork for future research, e.g., WSSS in the era of generative AIs.

Potential social impact Generative models, including SD, can be influenced by social biases in their training data, and our method of training segmentation models with the generated data may inherit the biases. Mitigating such biases is an ongoing research topic, and we have to be careful about deployment in socially sensitive or important usages. We used SD 1.4, which used the LAION dataset that was temporarily retracted for legal and safety concerns [37]. While Stable Diffusion 1.4 is kept published, we plan to replace it after a Stable Diffusion model trained on safe datasets is released.

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: CVPR. pp. 2209–2218 (2019) 12
2. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: ICLR (2022) 3
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. ICLR (2018) 11
4. Chen, C., Shu, K.: PromptDA: Label-guided data augmentation for prompt-based few-shot learners. arXiv preprint arXiv:2205.09229 (2022) 3
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018) 2, 10
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 10
7. Chen, Z., Wang, T., Wu, X., Hua, X.S., Zhang, H., Sun, Q.: Class re-activation maps for weakly-supervised semantic segmentation. In: CVPR. pp. 969–978 (2022) 5, 8, 12
8. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022) 2
9. Cho, J.H., Mall, U., Bala, K., Hariharan, B.: PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In: CVPR. pp. 16794–16804 (2021) 5, 13
10. 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: CVPR. pp. 3213–3223 (2016) 2, 9
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) 9
12. Dwibedi, D., Misra, I., Hebert, M.: Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: ICCV. pp. 1301–1310 (2017) 4
13. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. IJCV **111**, 98–136 (2015) 2, 9
14. Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., Tian, Y.: Scaling laws of synthetic images for model training... for now. In: CVPR. pp. 7382–7392 (2024) 1
15. Gao, S., Li, Z.Y., Yang, M.H., Cheng, M.M., Han, J., Torr, P.: Large-scale unsupervised semantic segmentation. IEEE TPAMI (2022) 9, 12
16. Ge, Y., Xu, J., Zhao, B.N., Joshi, N., Itti, L., Vineet, V.: Beyond generation: Harnessing text to image models for object detection and segmentation. arXiv preprint arXiv:2309.05956 (2023) 4
17. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014) 1
18. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. ICLR (2022) 13
19. Hammoud, H.A.A.K., Itani, H., Pizzati, F., Torr, P., Bibi, A., Ghanem, B.: SynthCLIP: Are we ready for a fully synthetic clip training? CVPRW (2024) 1

20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 10
21. Healthcare Intelligence Laboratory: SimpleCRF. https://github.com/HiLab-git/SimpleCRF (2017) 10
22. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2022) 2
23. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2021) 10
24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017) 11
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020) 1, 10
26. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: CyCADA: Cycle-consistent adversarial domain adaptation. In: ICML. pp. 1989–1998. Pmlr (2018) 4
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. ICLR (2022) 3, 9
28. Hugging Face: DreamBooth fine-tuning with LoRA (2023), https://huggingface.co/docs/peft/task_guides/dreambooth_lora 9, 11
29. Hugging Face: Stable-Diffusion-v1-4 (2023), https://huggingface.co/CompVis/stable-diffusion-v1-4 9
30. Isola, P.: Generative Models as Data++ (2023), hhttps://iplab.dmi.unict.it/icvss2023/Abstracts/Isola 1
31. Jo, S., Yu, I.J.: Puzzle-CAM: Improved localization via matching partial and full features. In: ICIP. pp. 639–643. IEEE (2021) 5
32. Jo, S., Yu, I.J., Kim, K.: MARS: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation. In: ICCV (2023) 12
33. Karazija, L., Laina, I., Vedaldi, A., Rupprecht, C.: Diffusion models for zero-shot open-vocabulary segmentation. arXiv preprint arXiv:2306.09316 (2023) 4
34. Kataoka, H., Hayamizu, R., Yamada, R., Nakashima, K., Takashima, S., Zhang, X., Martinez-Noriega, E.J., Inoue, N., Yokota, R.: Replacing labeled real-image datasets with auto-generated contours. In: CVPR. pp. 21232–21241 (2022) 4
35. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV (2023) 5, 12
36. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials. NeurIPS **24** (2011) 7
37. LAION.ai: Safety Review for LAION, 19 Dec, 2023 (2023), https://laion.ai/notes/laion-maintanence/ 14
38. Li, D., Ling, H., Kim, S.W., Kreis, K., Fidler, S., Torralba, A.: BigDatasetGAN: Synthesizing imagenet with pixel-wise annotations. In: CVPR. pp. 21330–21340 (2022) 3, 4, 11, 12
39. Li, Y., Duan, Y., Kuang, Z., Chen, Y., Zhang, W., Li, X.: Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In: AAAI. vol. 36, pp. 1447–1455 (2022) 5
40. Li, Y., Kuang, Z., Liu, L., Chen, Y., Zhang, W.: Pseudo-mask matters in weakly-supervised semantic segmentation. In: ICCV. pp. 6964–6973 (2021) 5, 8, 12

41. Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Open-vocabulary object segmentation with diffusion models. In: ICCV (2023) 4
42. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: CVPR. pp. 15305–15314 (2023) 12
43. Liu, S., Liu, K., Zhu, W., Shen, Y., Fernandez-Granda, C.: Adaptive early-learning correction for segmentation from noisy annotations. In: CVPR. pp. 2606–2616 (2022) 5
44. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021) 10
45. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015) 2
46. Ma, C., Yang, Y., Ju, C., Zhang, F., Liu, J., Wang, Y., Zhang, Y., Wang, Y.: DiffusionSeg: Adapting diffusion towards unsupervised object discovery. arXiv preprint arXiv:2303.09813 (2023) 3
47. MMSegmentation Contributors: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) 10
48. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation. NeurIPS (2023) 4, 8, 9, 11, 12
49. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML. pp. 16784–16804. PMLR (2022) 1
50. OpenAI: ChatGPT, Oct 16 version (2023), https://chat.openai.com/chat 9
51. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR. pp. 685–694 (2015) 2
52. Qi, L., Yang, L., Guo, W., Xu, Y., Du, B., Jampani, V., Yang, M.H.: UniGS: Unified representation for image generation and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6305–6315 (2024) 3
53. 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: ICML. pp. 8748–8763. PMLR (2021) 4
54. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. pp. 8821–8831. PMLR (2021) 1
55. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: DenseCLIP: Language-guided dense prediction with context-aware prompting. In: CVPR. pp. 18082–18091 (2022) 4, 5
56. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: ECCV. pp. 102–118. Springer (2016) 4
57. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) 1, 2
58. Rong, S., Tu, B., Wang, Z., Li, J.: Boundary-enhanced co-training for weakly supervised semantic segmentation. In: CVPR. pp. 19574–19584 (2023) 2, 5, 7, 11, 12

59. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: CVPR. pp. 3234–3243 (2016) 4
60. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22500–22510 (2023) 9
61. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR (2023) 1, 4
62. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. NeurIPS workshop (2021) 1, 9
63. Shin, G., Xie, W., Albanie, S.: ReCo: Retrieve and co-segment for zero-shot transfer. NeurIPS **35**, 33754–33767 (2022) 5, 13
64. Shinoda, R., Hayamizu, R., Nakashima, K., Inoue, N., Yokota, R., Kataoka, H.: SegRCDB: Semantic segmentation via formula-driven supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20054–20063 (2023) 4
65. Sun, W., Liu, Z., Zhang, Y., Zhong, Y., Barnes, N.: An alternative to WSSS? an empirical study of the segment anything model (SAM) on weakly-supervised semantic segmentation problems. arXiv preprint arXiv:2305.01586 (2023) 5, 12
66. Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., Ture, F.: What the DAAM: Interpreting stable diffusion using cross attention. Annual Meeting of the Association for Computational Linguistics (ACL) (2023) 2
67. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. arXiv preprint arXiv:2308.12469 (2023) 2, 13
68. Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., Isola, P.: Learning vision from models rivals learning vision from data. In: CVPR. pp. 15887–15898 (2024) 1
69. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: StableRep: Synthetic images from text-to-image models make strong visual representation learners. arXiv preprint arXiv:2306.00984 (2023) 1, 4
70. 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. arXiv preprint arXiv:2309.02773 (2023) 4
71. Wang, X., Ma, H., You, S.: Deep clustering for weakly-supervised semantic segmentation in autonomous driving scenes. Neurocomputing **381**, 20–28 (2020) 13
72. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: CVPR. pp. 12275–12284 (2020) 5
73. Wang, Y., Xu, C., Sun, Q., Hu, H., Tao, C., Geng, X., Jiang, D.: PromDA: Prompt-based data augmentation for low-resource NLU tasks. pp. 4242–4255 (2022) 3
74. Wei, J., Zou, K.: EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 6382–6388 (2019) 8
75. Wu, W., Dai, T., Huang, X., Ma, F., Xiao, J.: Image augmentation with controlled diffusion for weakly-supervised semantic segmentation. arXiv preprint arXiv:2310.09760 (2023) 5

76. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. ICCV (2023) 2, 4, 7, 8, 9, 10, 11, 12
77. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: CVPR. pp. 2955–2966 (2023) 3
78. Yang, L., Xu, X., Kang, B., Shi, Y., Zhao, H.: FreeMask: Synthetic images with dense annotations make stronger segmentation models. NeurIPS **36** (2024) 4
79. Yang, X., Gong, X.: Foundation model assisted weakly supervised semantic segmentation. In: WACV. pp. 523–532 (2024) 5
80. Yang, Z., Zhan, F., Liu, K., Xu, M., Lu, S.: Ai-generated images as data source: The dawn of synthetic era. arXiv preprint arXiv:2310.01830 (2023) 1
81. Yu, P., Xie, S., Ma, X., Zhu, Y., Wu, Y.N., Zhu, S.C.: Unsupervised foreground extraction via deep region competition. NeurIPS **34**, 14264–14279 (2021) 5
82. Zhang, D.J., Xu, M., Xue, C., Zhang, W., Han, X., Bai, S., Shou, M.Z.: Free-ATM: Exploring unsupervised learning on diffusion-generated images with free attention masks. arXiv preprint arXiv:2308.06739 (2023) 4
83. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543 (2023) 5
84. Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., Lu, J.: Unleashing text-to-image diffusion models for visual perception. ICCV (2023) 3
85. Zhao, Y., Ye, Q., Wu, W., Shen, C., Wan, F.: Generative prompt model for weakly supervised object localization. In: ICCV. pp. 6351–6361 (2023) 3
86. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016) 2, 5
87. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from CLIP. In: ECCV (2022) 2, 5, 13