ScribbleGen: Generative Data Augmentation Improves Scribble-supervised Semantic Segmentation

Jacob Schnell[1] * Jieke Wang[2] Lu Qi[2] Vincent Tao Hu[3] Meng Tang[2]
[1]University of Waterloo [2]University of California, Merced [3]CompVis Group, LMU Munich

Abstract

Recent advances in generative models, such as diffusion models, have made generating high-quality synthetic images widely accessible. Prior works have shown that training on synthetic images improves many perception tasks, such as image classification, object detection, and semantic segmentation. We are the first to explore generative data augmentations for scribble-supervised semantic segmentation. We propose ScribbleGen, a generative data augmentation method that leverages a ControlNet diffusion model conditioned on semantic scribbles to produce high-quality training data. However, naive implementations of generative data augmentations may inadvertently harm the performance of the downstream segmentor rather than improve it. We leverage classifier-free diffusion guidance to enforce class consistency and introduce encode ratios to trade off data diversity for data realism. Using the guidance scale and encode ratio, we can generate a spectrum of high-quality training images. We propose multiple augmentation schemes and find that these schemes significantly impact model performance, especially in the low-data regime. Our framework further reduces the gap between the performance of scribble-supervised segmentation and that of fully-supervised segmentation. We also show that our framework significantly improves segmentation performance on small datasets, even surpassing fully-supervised segmentation. The code is available at https://github.com/mengtang-lab/scribblegen.

1. Introduction

With the massive leaps forward in modern deep learning, machine learning model capacity has never been higher, with some models reaching billions of parameters [8, 11]. However, for many tasks, the size and complexity of datasets have not kept up with the explosion in model capacity. Since machine learning models perform with a large and rich training dataset, the question of scaling datasets to match model sizes is increasingly pressing. For tasks like fully-supervised semantic segmentation (FSSS), however, this is especially expensive due to the need for dense pixel-level annotations. These annotations must often also be produced by experts with domain-specific knowledge, exacerbating the costs of data labeling even further.

Weakly-supervised semantic segmentation (WSSS) seeks to reduce the requirement for dense annotations by using weak annotations. Such methods include scribble-supervised semantic segmentation, where only a fraction of

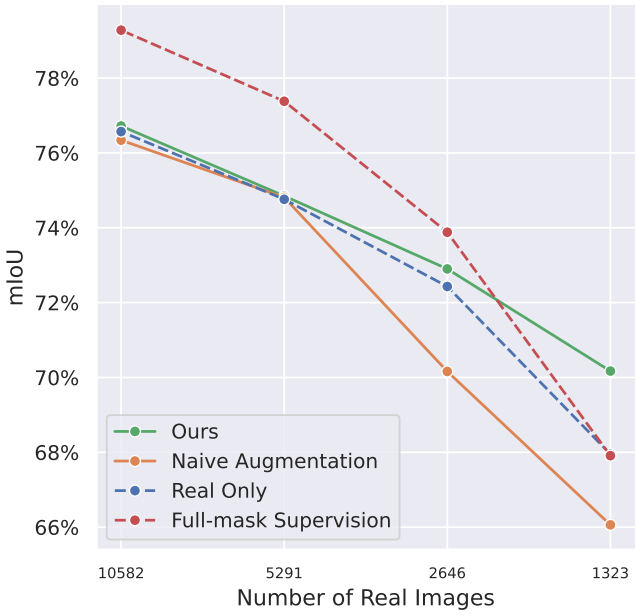

Figure 1. Segmentation model performance on PascalVOC and its subsets. All results other than full-mask supervision use scribble-supervision with ResNet-based RLoss [39] model. Naive data augmentation (i.e., fixed encode ratio $\lambda = 1.0$) harms model performance, especially in the low-data regime, while our augmentation scheme (Adaptive λ Sampling) improves performance.

*Work done during an internship at UC Merced.

Published at workshop on SyntaGen - Harnessing Generative Models for Synthetic Visual Datasets at IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

pixels along some lines (scribbles) are provided. However, these methods still lag behind fully-supervised alternatives regarding segmentation quality, with state-of-the-art methods still achieving 2-4% lower mIoU [26, 42] relative to fully-supervised models.

Another strategy is to produce synthetic training data using image-generative models. Prior works have shown that using Generative Adversarial Networks (GANs) to produce training data improves results in image classification [13] and semantic segmentation [3, 53], among other tasks. Diffusion models [17, 35, 37], a well-known type of generative models, have demonstrated strong performance in terms of controllability [32, 51] and fidelity [10, 33]. Several studies have successfully applied diffusion models to synthesize training data for image classification [1], object detection [52], and fully-supervised segmentation [43, 48]. This raises the question: *Can we also leverage the power of diffusion models to synthesize training data to further enhance the performance of scribble-supervised segmentation*?

In this work, we introduce ScribbleGen, a diffusion model conditioned on semantic scribbles to generate high-fidelity synthetic training images. Deep image-generative models such as diffusion models commonly used today, often require large datasets to produce high-quality images. This leads to a paradox where to upscale our training dataset, we need to already have access to a large training dataset. We address this problem by including a new parameter in the generative process, the encode ratio, which trades off image diversity for image photorealism.

Our contributions are summarized as follows:

- We are the first to leverage denoising diffusion models for generative data augmentation for scribble-supervised semantic segmentation. Our approach produces a spectrum of synthetic images conditioned on scribbles using different guidance scales and encode ratios.
- We provide detailed analyses and propose several schemes to combine synthetic and real data effectively for scribble-supervised semantic segmentation. We also identify the limitations of naive data augmentation schemes that can harm segmentation performance relative to not using synthetic training data at all.
- We achieve state-of-the-art results in scribble-supervised semantic segmentation, closing the gap between weakly-supervised and fully-supervised models as shown in Fig. 1. In particular, our framework significantly improves segmentation results in the low-data regime, where only a limited number of images are available.

2. Related work

Synthetic training data Numerous efforts have been dedicated to leveraging synthetic data for training perception models. IT-GAN [54] shows that GAN-generated samples can help classification models learn faster and improve performance. DatasetGAN [53], BigDatasetGAN [24], and HandsOff [50] employ GANs [4, 19] for jointly generating synthetic images and their corresponding labels for segmentation tasks.

Recent advances in diffusion models have brought notable stability during training, robust synthesis capabilities [10], and enhanced controllability [51]. As a result, there has been a significant shift towards the use of diffusion models for data synthesis, including for image classification [1, 20], object detection [52], instance segmentation [45], and semantic segmentation [25, 30, 43, 49]. For example, by fine-tuning an Imagen [33] model on ImageNet [9], [1] generates synthetic images from text prompts to improve the performance of image classification. Similarly, D3S [20] introduces a novel synthetic dataset specially designed to mitigate the foreground and background biases prevalent in real images. [30, 43, 44] jointly generate synthetic images and associated mask annotation, akin to DatasetGAN, using a StableDiffusion [32] image-generative model. GroundedDiffusion [25] further generates the triplet of image, mask, and texts to adapt the pre-trained diffusion model for open-vocabulary segmentation. FreeMask [49] utilizes FreestyleNet [48] to synthesize images conditioned on full mask annotations.

Our work diverges from these initiatives by focusing on sparse labels (e.g., scribbles) from real images as generative conditions, encouraging the creation of realistic and diverse synthetic images. While FreeMask [49] similarly conditions synthetic images on real data annotations, our method uses sparse rather than dense annotations, allowing for broader applications where dense labeling is expensive.

Guidance in Diffusion models Diffusion models excel in various tasks due to their controllability [51]. They're used to generate image content [17], image layout [18, 29, 32], audio content [28], human motion [40], etc. Guidance signals can also be incorporated to enhance image fidelity [10, 16] relative to unconditional generation. It has been shown that diffusion models can be guided by pretraining a noisy-data-based classifier, known as Classifier-guidance [10]. On the other hand, classifier-free guidance [16] removes the need for extra pretraining by randomly dropping out the guidance signal during training. We develop a framework that utilizes classifier-free guidance for generative data augmentation to improve scribble-based segmentation.

Weakly-supervised segmentation Weakly-supervised segmentation methods use weak annotations rather than full segmentation masks to train segmentation networks for images [6, 21, 26, 39, 42, 46] or point clouds [41]. Forms of weak annotations include points [2, 39, 42], scribbles [26, 39, 42], bounding boxes [22], image-level tags [6], and text [46]. These methods can be roughly

categorized into two groups. The first group proposes various unsupervised or semi-supervised losses such as entropy loss [5], CRF loss [39], and contrastive-learning losses [21]. The second group iteratively refines full-mask pseudo-labels [22, 27] during training to mimic full supervision. Many weakly-supervised approaches rely on class activation maps (CAMs) [5, 55] that gives localization cues from classification networks. Our generative data augmentation approach complements any existing weakly-supervised segmentation methods, as we show improved performance of several methods with our synthetic data.

Weak annotations can also be provided as input for segmentation networks at test time for interactive segmentation [23, 47]. For example, Segment Anything [23] allows prompts including clicks, bounding boxes, masks, or text. While Segment Anything [23] provides many masks in a semi-automatic way for training interactive segmentation, we focus on synthetic image synthesis for training weakly-supervised segmentation.

3. Method

In this section, we describe our method of generative data augmentation for weakly-supervised semantic segmentation outlined in Fig. 2. First, in Sec. 3.1, we provide a background on sampling from diffusion models. Then, in Sec. 3.2, we introduce a variant of ControlNet [51] conditioned on scribble labels and text prompts. We further discuss how to achieve semantically consistent images and trade off diversity and photorealism through guided diffusion and encode ratio in Sec. 3.3 and Sec. 3.4, respectively. Sec. 3.5 proposes several schemes to effectively combine synthetic and real images for training segmentation networks.

3.1. Background

Diffusion models. Diffusion models [17, 35, 37] learn to reverse a forward process that gradually adds noise to an image $\mathbf{x}_{\text{ref}}$ until the original signal is fully diminished. After training, following the reverse process allows us to sample an image $\mathbf{x}_0$ given noise $\mathbf{x}_T \sim \mathcal{N}(0, I)$. Learning this reverse process reduces to learning a denoiser $\boldsymbol{\epsilon}_\theta$ that recovers the original image from a noisy image $\mathbf{x}_t$ as

$$\mathbf{x}_{\text{ref}} \approx f_\theta(\mathbf{x}_t, t) := (\mathbf{x}_t - (1 - \overline{\alpha}_t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t))/\sqrt{\overline{\alpha}_t}. \quad (1)$$

To get high-quality samples, the standard diffusion model sampling process [17] requires many (often $T = 1,000$) neural function evaluations. Using a non-Markovian forward process, Denoising Diffusion Implicit Model (DDIM) [36] samplers forego several intermediate function calls, accelerating sampling. Let τ be an increasing subsequence of $[T, \dots, 1]$ and define the DDIM forward process for some stochasticity parameter $\sigma \in \mathbb{R}^T_{\geq 0}$ as

$$q_\sigma(\mathbf{x}_{\tau_{i-1}}|\mathbf{x}_{\tau_i}, \mathbf{x}_0) = \mathcal{N}\Bigg(\sqrt{\alpha_{\tau_{i-1}}}\mathbf{x}_0 + \sqrt{1 - \alpha_{\tau_{i-1}} - \sigma^2_{\tau_i}} \cdot \frac{\mathbf{x}_{\tau_i} - \sqrt{\alpha_{\tau_i}}\mathbf{x}_0}{\sqrt{1 - \alpha_{\tau_i}}},\ \sigma^2_{\tau_i} I\Bigg). \quad (2)$$

We can then sample from the generative process using the abovementioned forward process. In particular, using $f_\theta(x_t, t)$ as defined in Eq. (1) we can sample $\mathbf{x}_{\tau_{i-1}}$ from $\mathbf{x}_{\tau_i}$ by

$$p^{(\tau_i)}_\theta(\mathbf{x}_{\tau_{i-1}}|\mathbf{x}_{\tau_i}) = \begin{cases} \mathcal{N}(f_\theta(\mathbf{x}_{\tau_i}, \tau_i),\ \sigma^2_{\tau_i} I) & \text{if } i = 1 \\ q_\sigma(\mathbf{x}_{\tau_{i-1}}|\mathbf{x}_{\tau_i}, f_\theta(\mathbf{x}_{\tau_i}, \tau_i)) & \text{if } i > 1 \end{cases} \quad (3)$$

We slightly abuse notation here and define $\tau_0 = 0$ so that when $i = 1$, we sample the denoised image $\mathbf{x}_0$.

Classifier-free guidance. To trade off mode coverage and sample fidelity in a conditional diffusion model, [10] proposes to guide the image generation process using the gradients of a classifier, with the additional cost of training the classifier on noisy images. To address this drawback, classifier-free guidance [16] does not require any classifier. They obtain a conditional and unconditional network combination in a single model by randomly dropping the guidance signal $\mathbf{c}$ during training. After training, it empowers the model with progressive control over the degree of alignment between the guidance signal and the sample by varying the guidance scale w when a larger w leads to greater alignment with the guidance signal:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t;\ \mathbf{c}, w) = (1 + w)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t;\ \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t). \quad (4)$$

3.2. Scribble-conditioned Image Synthesis

We consider a semantic synthesis approach to generating our synthetic training data. The synthetic training data is generated conditioned on real segmentation labels from the training dataset. We leverage a typical denoising diffusion model, ControlNet [51], to achieve image synthesis conditioned on the segmentation scribbles. Our model is trained using the usual DDPM [17] object: given a noisy image $\mathbf{x}_t$ (in reality $\mathbf{x}_t$ is a latent representation as in [32], but we omit this detail for brevity) and conditioning input $\mathbf{c}$ it predicts the added noise $\boldsymbol{\epsilon}$. Our segmentation scribbles on which the model is conditioned are represented as RGB images in $\mathbb{R}^{h \times w \times 3}$ with different colors for every class, though we explore other representations in Sec. 4.3.

Finally, we note that it is difficult for the ControlNet model to produce semantically consistent images with the given scribble labels. We hypothesize that this is due to

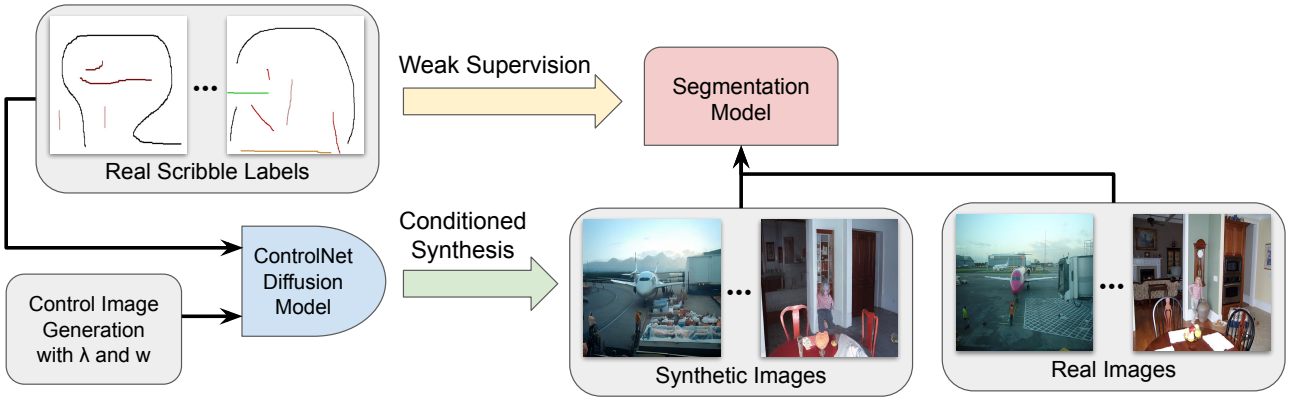

Figure 2. Given a limited number of real scribbles, we pretrain a ControlNet-based diffusion model for high-fidelity synthesis of images conditioned on scribbles. We can control the image synthesis with the encode ratio λ and the guidance scale w. These image-scribble pairs can then be smoothly integrated into the training of scribble-based semantic segmentation.

the difficulty of encoding class information in RGB images, especially in the early stages of training. Therefore, we supplement our model with text prompts that include all the classes within the image. Adding these prompts significantly improves image class consistency and leads to higher-quality images relative to an unchanging default prompt. We explore the effect of this prompt in Sec. 4.3.

Our ControlNet training objective is thus

$$\mathcal{L}_{\text{CN}}(\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{x}_{\text{ref}},\mathbf{c}_s,\mathbf{c}_t),t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{c}_s, \mathbf{c}_t)\|_2^2\right], \quad (5)$$

where $(\mathbf{x}_{\text{ref}}, \mathbf{c}_s, \mathbf{c}_t)$ is the triplet of the original (unnoised) image, the conditioning scribble label, and the conditioning text prompt and $\boldsymbol{\epsilon}_\theta$ is our ControlNet diffusion model.

3.3. Classifier-free Scribble Guidance

We leverage diffusion guidance to further improve semantic consistency between the generated synthetic image and conditional input. Following the proposals from Classifier-free Guided Diffusion [16], we randomly drop out 10% of all conditioning scribble inputs $\mathbf{c}_s$, replacing them with a randomly initialized, learned embedding $\tilde{\mathbf{c}}$, when training the ControlNet model. By modifying Eq. 4, we arrive at a new guided noise prediction function:

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t; \mathbf{c}_s, \mathbf{c}_t, w) = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t; \mathbf{c}_s, \mathbf{c}_t) - w\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t; \tilde{\mathbf{c}}). \quad (6)$$

While ControlNet uses a pre-trained Stable-Diffusion model [32], which is trained conditionally and unconditionally, scribble drop-out during training can be viewed as fine-tuning the unconditional diffusion model to our dataset. We have found that the guidance scale, w, can significantly impact the quality of generated images, especially with respect to the fine-grain details of the produced image. We further ablate this hyperparameter's impact in Sec 4.3.

3.4. Control Image Diversity via Encode Ratio

The vanilla diffusion model denoises sampled Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, I)$ iteratively until $\mathbf{x}_0$ at inference time. In practice, synthetic images generated this way may be unrealistic, particularly when training data is limited for our scribble-conditioned diffusion model. To improve photorealism at the cost of diversity, we propose another forward diffusion process parameter, the encode ratio $\lambda \in (0, 1]$. Specifically, we perform $\lambda \cdot T$ noise-adding forward diffusion steps to the input images and, during inference, denoise $\mathbf{x}_{\lambda T}$ iteratively until $\mathbf{x}_0$. Thus, for $\lambda = 1$, there is no change, but for small choices of λ, there is less noise added to the image $\mathbf{x}_0$. As $\lambda \to 0$, the sampled image will become increasingly similar to the original $\mathbf{x}_{\text{ref}}$. Therefore, a whole spectrum of synthetic images with varying levels of similarity to the reference image can be achieved by varying our choice of λ. We outline our sampling algorithm, which combines the accelerated DDIM sampling from Sec. 3.1, the scribble guidance from Sec. 3.3, and the encode ratio from Sec 3.4 in Algorithm 1. Fig. 3 shows synthetic images generated with varying guidance scales and encode ratios.

3.5. Combine synthetic images with real images

Generative data augmentation can, in principle, produce an infinite amount of synthetic images. However, naively combining real and synthetic images can harm rather than benefit weakly-supervised segmentation models, as we have observed. In particular, it is not clear which choices of the guidance scale w and encode ratio λ are optimal. We choose the optimal guidance scale, as determined in Sec. 4.3. For encode ratio λ, we propose and systematically evaluate two strategies for combining synthetic with real images.

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ denote the set of all real im-

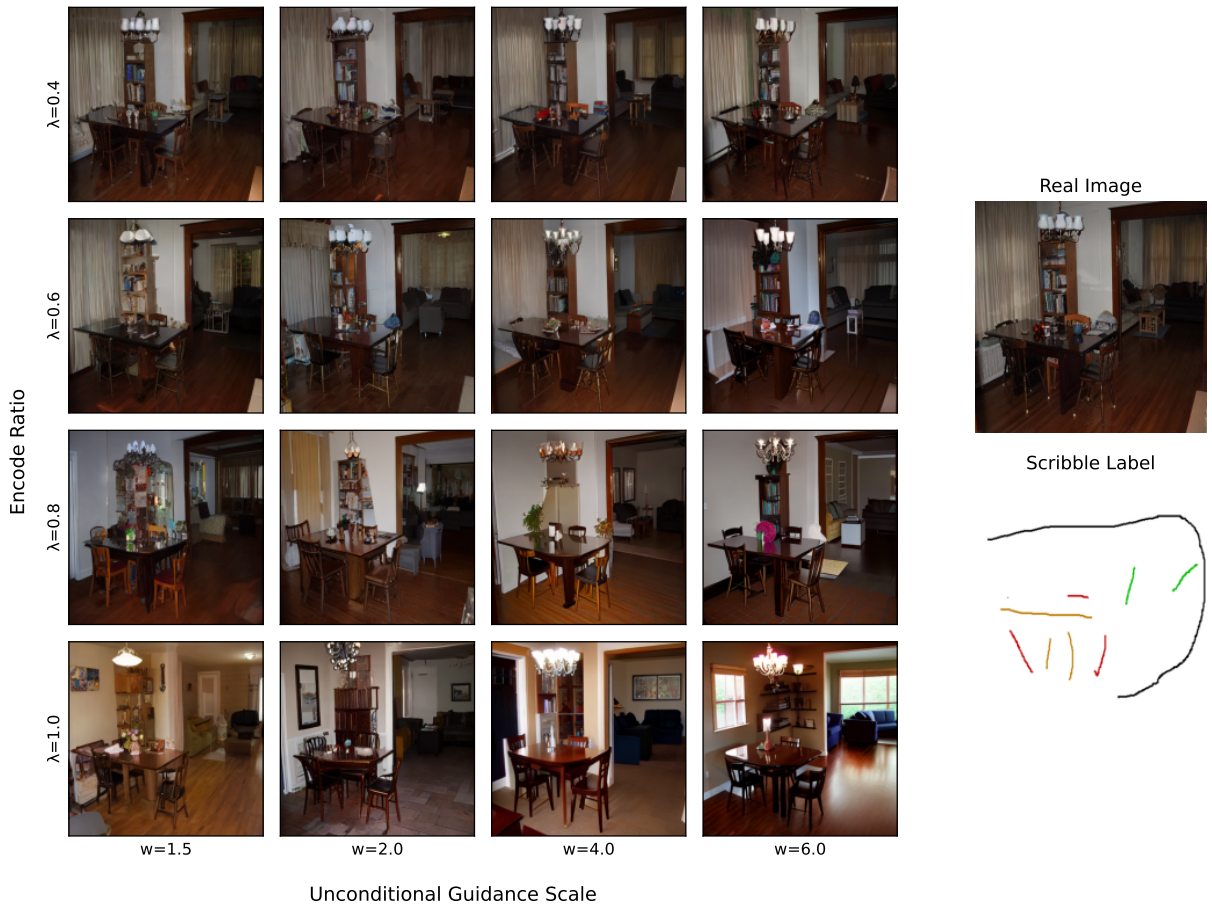

Figure 3. Left: Our sampled synthetic images conditioned on the ground-truth scribble. By sampling using different guidance scales and encode ratios we are able to generate a whole spectrum of realistic synthetic training images. Right: The ground-truth real image and corresponding scribble label.

ages and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ denote the set of all (scribble) labels. Then we produce a set of synthetic images $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n\}$ where $\hat{\mathbf{x}}_i = \mathrm{DM}_\theta(\mathbf{y}_i, \mathbf{c}_i; w, \lambda)$ is the output of our trained diffusion model, DM_θ, conditioned on the scribble $\mathbf{y}_i$ and prompt-condition $\mathbf{c}_i$, given guidance scale w and encode ratio λ. We may then produce a new, augmented dataset $\mathcal{X}' = \mathrm{concat}(\mathcal{X}, \hat{\mathcal{X}})$ and $\mathcal{Y}' = \mathrm{concat}(\mathcal{Y}, \mathcal{Y})$. Note this means each label, $\mathbf{y}_i$, appears twice in our dataset, once for the real image $\mathbf{x}_i$ and once for the synthetic image $\hat{\mathbf{x}}_i$.

- **Fixed encode ratio** λ: We choose a fixed encode ratio which gives a fixed synthetic dataset $\hat{\mathcal{X}}$. Using the default value of $\lambda = 1$ yields the most diverse synthetic images with possibly inferior image fidelity. We find the optimal λ that gives the best segmentation in our experiments.
- **Adaptive encode ratio** λ: To avoid hyper-parameter search, we also propose an adaptive scheme for choosing λ. We gradually increase the encode ratio λ while training downstream segmentation networks, similar to curriculum learning. Initially, synthetic images used for training are similar to real images, which are considered an easier curriculum to learn. Synthetic images diverge increasingly from the real images as training progresses. For this case, the synthetic dataset is formed at epoch e as $\hat{\mathcal{X}} = \{\hat{\mathbf{x}}_{1,\lambda_e}, \ldots, \hat{\mathbf{x}}_{1,\lambda_e}\}$ where we follow the encode ratio schedule $[\lambda_1, \ldots, \lambda_E] \in \Lambda^E$ where E is the number of training epochs.

4. Experiments

Sec. 4.1 summarize our main results that show improvements on several scribble-supervised segmentation methods using our generative data augmentation. In Sec. 4.2, we further explore the challenging scenario with limited number of real images. We show that naive implementations of generative data augmentation may harm the performance, whereas our data augmentation scheme improves. Sec. 4.3 gives an ablation study on guidance scale and encode ratio, two critical degrees of freedom for our image synthesis.

Dataset and Implementation Details We report results on the standard PASCAL VOC12 segmentation dataset

Algorithm 1: Conditional DDIM sampling with guidance scale w and encode ratio λ

Require: q_σ: forward process
Require: $\mathbf{x}_{\text{ref}}$: a reference image
Require: $w \geq 0$: guidance scale
Require: $\lambda \in [0, 1]$: encode ratio
Require: $N \in \{1, \dots, T\}$: number of reverse diffusion process steps
Require: c_t and c_s: text prompt and scribble conditioning

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$
$\tau = [\lfloor \frac{\lambda T}{N} n \rfloor : 0 \leq n \leq N]$
▷ Note if $\lambda = 1$ then $\mathbf{x}_{\tau_N} \sim \mathcal{N}(0, I)$
$\mathbf{x}_{\tau_N} = \sqrt{\alpha_{\tau_N}}\mathbf{x}_{\text{ref}} + \sqrt{1 - \alpha_{\tau_N}}\boldsymbol{\epsilon}$
for $i = N$ *to* 1 **do**
  ▷ Predict added noise using diffusion guidance (6)
  $\tilde{\boldsymbol{\epsilon}}_{\tau_i} = (1 + w)\boldsymbol{\epsilon}_\theta^{(\tau_i)}(\mathbf{x}_{\tau_i}, c_t, c_s) - w\boldsymbol{\epsilon}_\theta^{(\tau_i)}(\mathbf{x}_{\tau_i}, c_t, \tilde{c}_s)$
  ▷ Accelerated DDIM sampling (3)
  $\hat{\mathbf{x}}_0 = (\mathbf{x}_{\tau_i} - \sqrt{1 - \alpha_{\tau_i}} \cdot \tilde{\boldsymbol{\epsilon}}_{\tau_i})/\sqrt{\alpha_{\tau_i}}$
  if $i = 1$ **then**
    $\mathbf{x}_0 \sim \mathcal{N}(\hat{\mathbf{x}}_0, \sigma_{\tau_1}^2 I)$
  else
    $\mathbf{x}_{\tau_{i-1}} \sim q_\sigma(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}, \hat{\mathbf{x}}_0)$
  end
end for
return $\mathbf{x}_0$

which contains 10 582 images for training and 1 449 images for validation. We utilize scribbles from ScribbleSup dataset [27] with only 3% pixels labeled on average.

For image synthesis, we use a latent diffusion model [32] with a downsampling rate of $f = 8$, so that an input image of size 512×512 is downsampled to 64×64. We use Stable Diffusion 1.5 as the backbone for ControlNet [51] and finetune ControlNet for 200 epochs with a batch size of 16 using two A100 80GB GPUs. We set $T = 1000$ discrete timesteps for ControlNet and use a linear learning rate scheduler from an initial rate of 10^{-4} during training. For scribble conditioning, we randomly dropout 10% of scribbles, replacing them with a learned embedding of the same size. Scribble labels are represented as RBG images in $\{1, \dots, 255\}^{512 \times 512 \times 3}$. We also provide the text prompt "a high-quality, detailed, and professional image of [list of classes]" as suggested in [51]. We provide visualizations of our synthetic dataset in the supplementary material.

Evaluation metric. We evaluate both the diversity and fidelity of the generated images by the Fréchet Inception Distance (FID) [15], as it is the *de facto* metric for the evaluation of generative methods, e.g., [4, 10, 19, 33]. It provides a symmetric measure of the distance between two distributions in the feature space of Inception-V3 [38]. We use FID as our primary metric for the sampling quality. We realize, however, that FID should not be the only metric for evaluating the downstream impact of synthetic data for training segmentation networks. Hence, we also report segmentation results trained with synthetic data only to evaluate synthetic data, similar to the Classification Accuracy Score (CAS) proposed by [31] but for semantic segmentation. We report the standard mean Intersection Over Union (mIOU) metric for segmentation results.

4.1. Generative data augmentation improves scribble-supervised semantic segmentation

For our experiments, we consider two methods of weakly-supervised semantic segmentation, including simple regularized losses (RLoss) [39] and the current state-of-the-art in scribble-supervised segmentation, Adaptive Gaussian Mixture Models (AGMM) [42]. For both methods, we jointly train them on the original training set and our augmented training set. Both methods also follow a polynomial learning rate scheduler. The sampling of synthetic training images is outlined in Sec. 3.5. Table 1 shows improved results using generative data augmentation for both RLoss and AGMM. Our method with synthetic data further reduces the gap between weakly-supervised and fully-supervised segmentation. We show visualizations of our segmentation results with and without using our generative data augmentation in Fig. 6. We also include further visualizations in the supplementary material.

4.2. Low-data Regime Results

For the low-data regime, we only consider the RLoss method due to its simplicity and speed to train. We consider three different reduced datasets with 50%, 25%, and 12.5% of all training images used, respectively. For each of these cases, we train a ControlNet diffusion model on the limited dataset (following the same experimental setup described at the start of Sec. 4) and sample synthetic images as usual. The results of training RLoss on each of the reduced datasets for our different proposed augmentation schemes are reported in Fig. 1.

We notice that the naive data augmentation fails to help in all of our reduced datasets and instead reduces model performance in all but the 50% case. Conversely, our proposed *Adaptive λ Sampling* improves or matches performance for all four datasets. We hypothesize this is due to the lack of training images required to ensure high-quality generation from our diffusion model. This hypothesis is confirmed by the significantly higher FID scores for synthetic datasets generated with limited training data reported in Fig. 5 middle. We also confirm this hypothesis qualitatively in Fig. 4, where we observe that fully synthetic images deteriorate in quality as the number of training images decreases. How-

Method	Network	Supervision	Synthetic Data	Augmentation Scheme	mIoU (%)
(1) *DeeplabV3+ [7]	MobileNet [34]	Full mask		–	72.1
(2) *DeeplabV3+ [7]	ResNet101 [14]	Full mask		–	79.3
RLoss [39]	(1)	Scribble		–	68.4
RLoss [39]	(1)	Scribble	✓	Fixed $\lambda = 1.0$	69.4 (+1.0)
RLoss [39]	(1)	Scribble	✓	Fixed $\lambda = 0.5$	**70.0** (+1.6)
RLoss [39]	(2)	Scribble		–	76.6
RLoss [39]	(2)	Scribble	✓	Fixed $\lambda = 1.0$	76.1 (-0.5)
RLoss [39]	(2)	Scribble	✓	Fixed $\lambda = 0.7$	**77.0** (+0.4)
AGMM [42]	(2)	Scribble		–	76.4
*AGMM [42]	(2)	Scribble		–	78.1
AGMM [42]	(2)	Scribble	✓	Fixed $\lambda = 1.0$	78.0 (-0.1)
AGMM [42]	(2)	Scribble	✓	Adaptive λ	78.7 (+0.6)
AGMM [42]	(2)	Scribble	✓	Fixed $\lambda = 0.4$	**78.9** (+0.8)

Table 1. Generative data augmentation improves scribble-supervised semantic segmentation methods including RLoss [39] and AGMM [42] on PascalVOC [12]. The best results are shown in **bold**. Numbers in parenthesis are relative improvement / decrease in comparison to the baseline without synthetic data. Note that * AGMM is our re-implementation which gives better results than reported [42].

ever, we can stabilize this deterioration by decreasing the encode ratio λ to improve image realism. Using our *Adaptive λ sampling*, the most synthetic (and thus lowest quality) images cannot impact model training as significantly due to the reduced learning from our scheduler.

4.3. Ablation Studies

Guidance Scale We report the FID scores of our fully synthetic ($\lambda = 1$) datasets as generated by our model trained on all PascalVOC training images in Fig. 5 left. This ablation study is how we decided to use $w = 2$ for all other experiments, as it yields optimal FID. We include further visualizations of the impact of the guidance scale on image synthesis in our supplementary material.

Encode Ratio We report the FID scores of our diffusion models trained with a variable number of images as a function of the encode ratio λ in Fig. 5 middle. We observe that the FID increases significantly as the number of training images decreases. However, we can reduce the effect of limited training data by decreasing the encode ratio to promote image realism. This effect is most pronounced for the 1323 image-trained diffusion model, where we reduce the FID score by over 30 points by lowering the encode ratio.

We also evaluate segmentation model performance on synthetic data of varying encode ratios and report the final mIoU in Fig. 5 right. For these experiments, we train segmentation models training using the Fixed λ data augmentation proposed in Sec. 3.5 and training exclusively on synthetic training data (i.e., $\mathcal{X}' = \hat{\mathcal{X}}$), akin to CAS [31]. We observe that the impact of varying the encode ratio λ is limited in the data augmentation case but much more significant for the synthetic-only case. We suppose that for the

Figure 4. Synthetic images sampled from diffusion models with different numbers of training images and encode ratios λ.

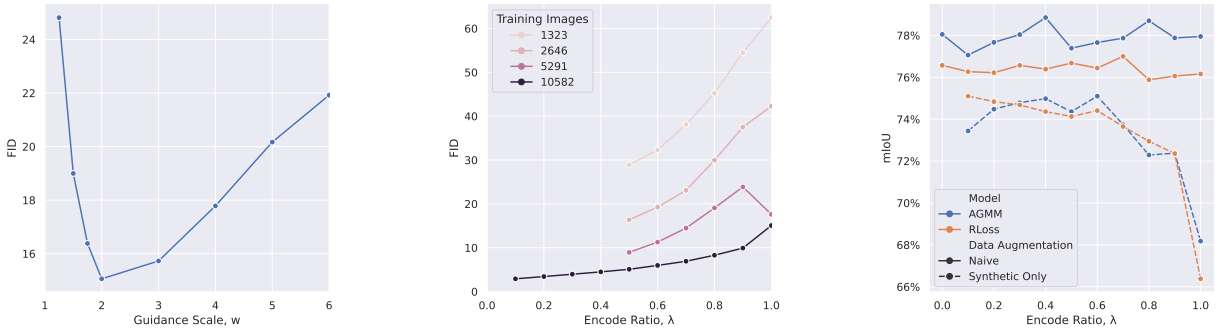

Figure 5. Left: The FID of our full training dataset when generated with different classifier-free guidance scales. Results are reported for ControlNet trained all 10582 images. Middle: The FID of our training dataset when generated with different encode ratios. Results are reported for four ControlNet models trained on a different number of images. Right: The mIoU of a downstream segmentation model when trained on datasets of varying encode ratios. Note $\lambda = 0.0$ corresponds to training on real images only. Results are reported for training on both naive data augmentation and only on synthetic images. In both cases, we use all 10582 images for training.

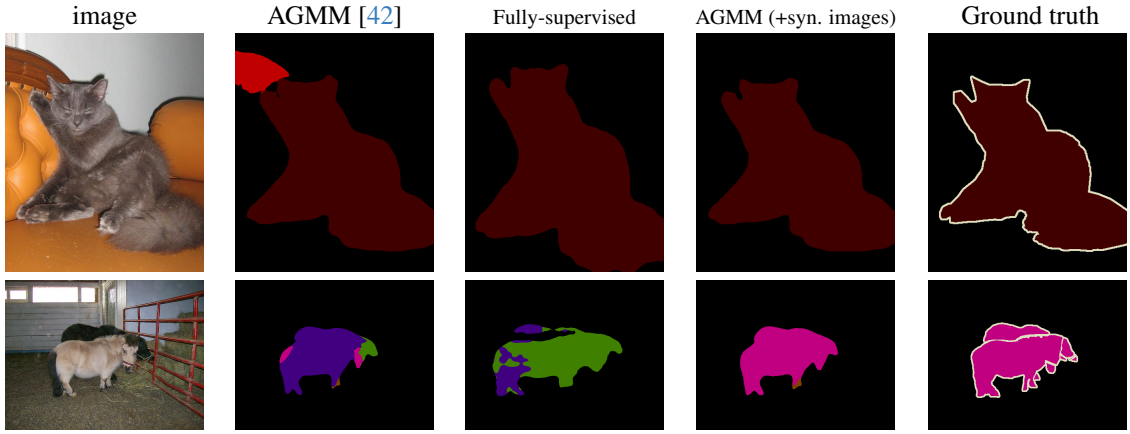

Figure 6. Qualitative results on PASCAL dataset. Our generative data augmentation method improves scribble-supervised semantic segmentation methods such as AGMM [42].

synthetic-only case, the quality of the synthetic images is more important, so decreasing the encode ratio to improve data realism matters more than data diversity. We include further visualizations of the impact of the encode ratio on image synthesis in our supplementary material.

Conditioning Input We also ablate modifying the conditioning input to ControlNet. We try representing scribble labels as one-hot embeddings in $\{0, 1\}^{h \times w \times C}$ where there are C total classes. Using these one-hot embeddings, we obtained a higher FID by 4.4 points relative to RGB embeddings, but we found no improvement in mIoU results using our Fixed λ augmentation scheme. We also try using text prompts that don't include the classes in the image. Using unchanging prompts (i.e., "a high-quality, detailed, and professional image") yields lower FID by 3.1 points relative to prompts that include the classes in the image and 1.9% lower mIoU using our Fixed λ augmentation scheme.

5. Conclusion and Future Work

We propose leveraging diffusion models conditioned on scribbles to produce high-quality synthetic training data for scribble-supervised semantic segmentation. We advocate the use of classifier-free guided diffusion and introduce the encode ratio to control the generative process, allowing us to generate a spectrum of images. We report state-of-the-art performance on scribble-supervised semantic segmentation with our generative data augmentation.

In the future, it will be interesting to train generative models for open-vocabulary image synthesis conditioned on sparse annotations. Our generative data augmentation has the potential to improve semi-supervised segmentation. We are also interested in end-to-end training of generative data augmentation and perception models, as metrics like FID are loosely related to perception performances.

Acknowledgement The authors would like to thank Prof. Yuri Boykov, Prof. Olga Veksler, and Prof. Ming-Hsuan Yang for their helpful discussion and comments that improved the quality of this work.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[2] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 2

[3] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger N. Gunn, Alexander Hammers, David Alexander Dickie, Maria del C. Valdés Hernández, Joanna M. Wardlaw, and Daniel Rueckert. GAN augmentation: Augmenting training data using generative adversarial networks. *CoRR*, abs/1810.10863, 2018. 2

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 2, 6

[5] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization. *CoRR*, abs/2008.01201, 2020. 3

[6] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020. 2

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. In *PAMI*, 2018. 7

[8] Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, and Neil Houlsby. Scaling vision transformers to 22 billion parameters, 2023. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 2

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3, 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 1

[12] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. In *IJCV*, 2010. 7

[13] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 289–293, 2018. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop*, 2021. 2, 3, 4

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[18] Tao Hu, David W Zhang, Yuki M. Asano, Gertjan J. Burghouts, and Cees G.M. Snoek. Self-guided diffusion models. In *CVPR*, 2023. 2

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 6

[20] Priyatham Kattakinda, Alexander Levine, and Soheil Feizi. Invariant learning via diffusion dreamed distribution shifts, 2022. 2

[21] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. In *International Conference on Learning Representations*, 2021. 2, 3

[22] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017. 2, 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[24] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdatasetgan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21330–21340, 2022. 2

[25] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7667–7676, 2023. 2

[26] Zhiyuan Liang, Tiancai Wang, Xiangyu Zhang, Jian Sun, and Jianbing Shen. Tree energy loss: Towards sparsely annotated semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16907–16916, 2022. 2

[27] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 3, 6

[28] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023. 2

[29] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 2

[30] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic dataset generation for pixel-level semantic segmentation, 2023. 2

[31] Suman V. Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *CoRR*, abs/1905.10887, 2019. 6, 7

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3, 4, 6

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2, 6

[34] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *CoRR*, abs/1801.04381, 2018. 7

[35] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2, 3

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3

[37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 2, 3

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6

[39] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018. 1, 2, 3, 6, 7

[40] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 2

[41] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2697–2707, 2022. 2

[42] Linshan Wu, Zhun Zhong, Leyuan Fang, Xingxin He, Qiang Liu, Jiayi Ma, and Hao Chen. Sparsely annotated semantic segmentation with adaptive gaussian mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15454–15464, 2023. 2, 6, 7, 8

[43] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models, 2023. 2

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

[45] Jiahao Xie, Wei Li, Xiangtai Li, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Mosaicfusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *arXiv preprint arXiv:2309.13042*, 2023. 2

[46] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 2

[47] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep grabcut for object selection. *arXiv preprint arXiv:1707.00243*, 2017. 3

[48] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis, 2023. 2

[49] Lihe Yang, Xiaogang Xu, Bingyi Kang, Yinghuan Shi, and Hengshuang Zhao. Freemask: Synthetic images with dense annotations make stronger segmentation models. In *NeurIPS*, 2023. 2

[50] Zuhao Yang, Fangneng Zhan, Kunhao Liu, Muyu Xu, and Shijian Lu. Ai-generated images as data source: The dawn of synthetic era. *arXiv preprint arXiv:2310.01830*, 2023. 2

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 6

[52] Manlin Zhang, Jie Wu, Yuxi Ren, Ming Li, Jie Qin, Xuefeng Xiao, Wei Liu, Rui Wang, Min Zheng, and Andy J Ma. Diffusionengine: Diffusion model is scalable data engine for object detection. *arXiv preprint arXiv:2309.03893*, 2023. 2

[53] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 2

[54] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan, 2022. 2

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

ScribbleGen: Generative Data Augmentation Improves Scribble-supervised Semantic Segmentation

Supplementary Material

Training Images	FID
10582	43.3
5291	53.7
2646	57.1
1323	58.0

Table 2. FID reported on the validation set. Synthetic images are from our ControlNet model with a varying number of training images. Synthetic images are conditioned on scribbles from the validation set, previously unseen to our model.

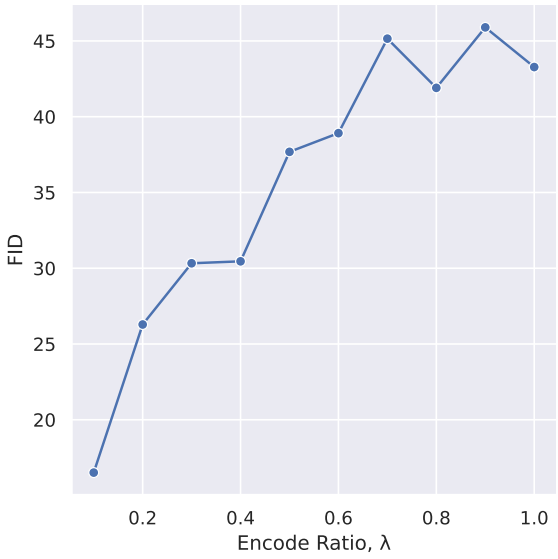

Figure 7. FID reported on the validation set. Synthetic images are from our ControlNet model trained on all of PascalVOC. Images are synthesized conditioned on scribbles from the validation set with varying encode ratios.

6. Validation Set FID Results

In this section, we report the FID of our synthetic images on the validation set. To achieve this, we provide scribbles from the validation set of PascalVOC as conditioning input to our trained ControlNet models. Since the ControlNet models were not trained with data from the validation set, these are previously unseen scribbles. In Table 2, we report the impact of the number of training images of the ControlNet model on validation FID. In Fig. 7, we report the impact of the encode ratio on validation FID.

7. Additional Qualitative Results

In this section, we include additional qualitative results of our method. We provide additional samples of our training data in Fig. 8 and samples from previously unseen scribbles from the validation set in Fig. 9. We also provide visualizations of the effect of the guidance scale on synthesis in Fig. 10 and the effect of the encode ratio in Fig. 11. The effect of the number of training images on synthesis is demonstrated in Fig. 12. Finally, we provide further visualizations of segmentation results in Fig. 13.

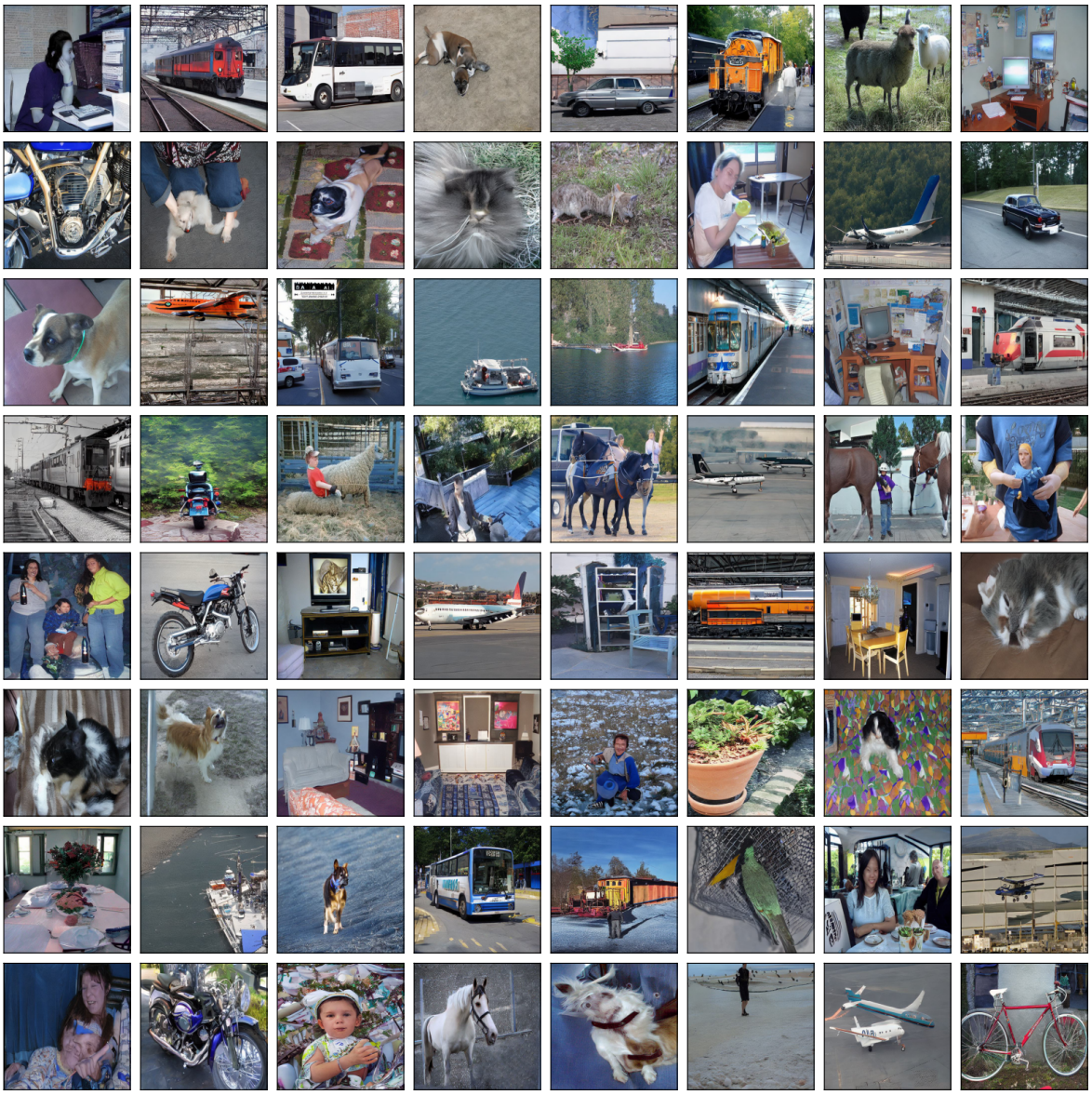

Figure 8. Synthetic training images sampled from a ControlNet model trained on all of scribble-supervised PascalVOC. All images are sampled using guidance scaled $w = 2.0$ and encode ratio $\lambda = 1.0$.

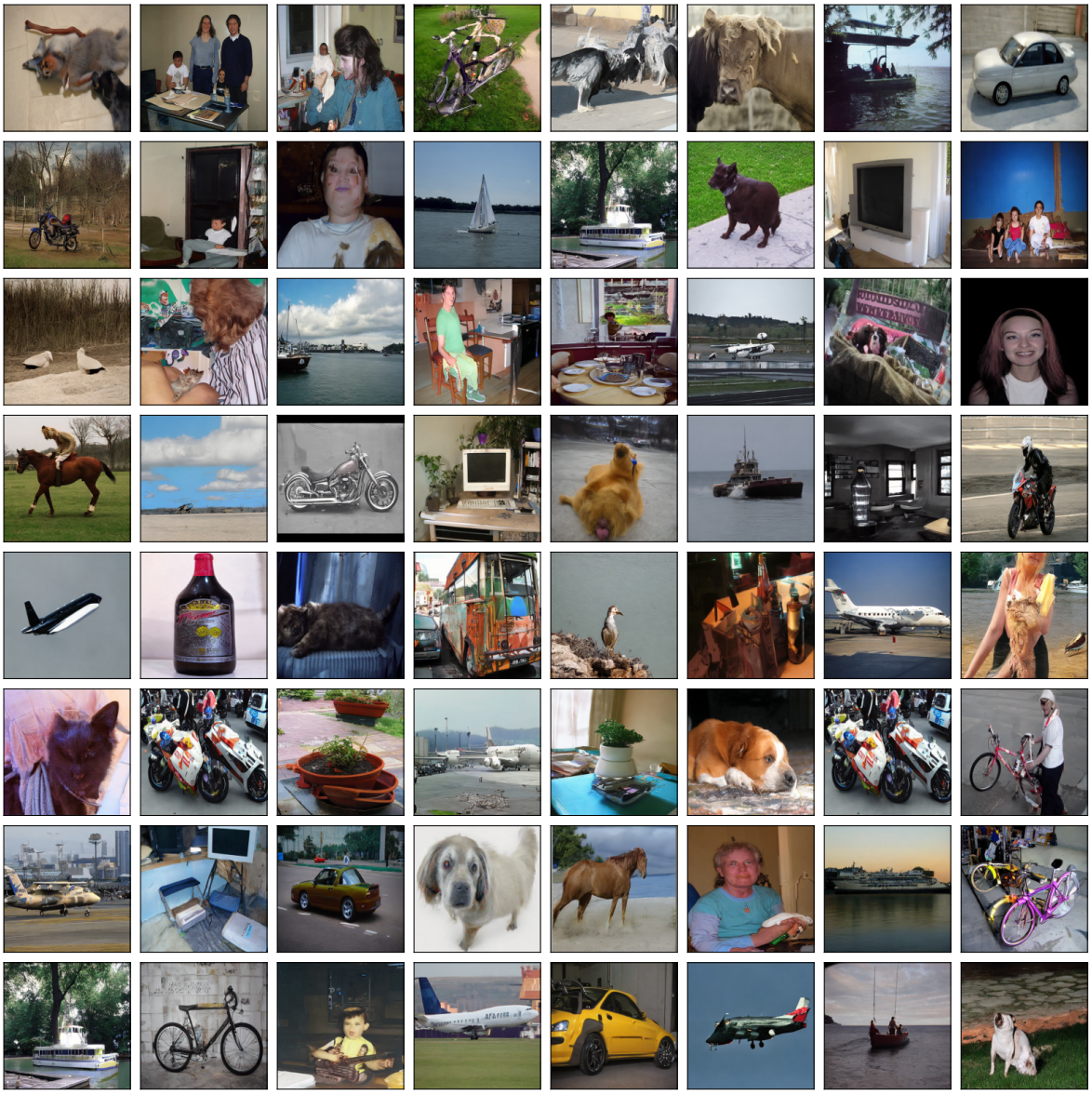

Figure 9. Synthetic validation images sampled from a ControlNet model trained on all of scribble-supervised PascalVOC. Images are synthesized conditioned on scribbles from the validation set, which the ControlNet model has not been trained on. All images are sampled using guidance scaled $w = 2.0$ and encode ratio $\lambda = 1.0$.

Figure 10. Synthetic training images sampled from a ControlNet model trained on all of scribble-supervised PascalVOC. We vary the guidance scale but keep the encode ratio $\lambda = 1.0$ constant to see the effect of the guidance scale on synthesis.

Figure 11. Synthetic training images sampled from a ControlNet model trained on all of scribble-supervised PascalVOC. We vary the encode ratio but keep the guidance scale $w = 2.0$ constant to see the effect of the encode ratio on synthesis.

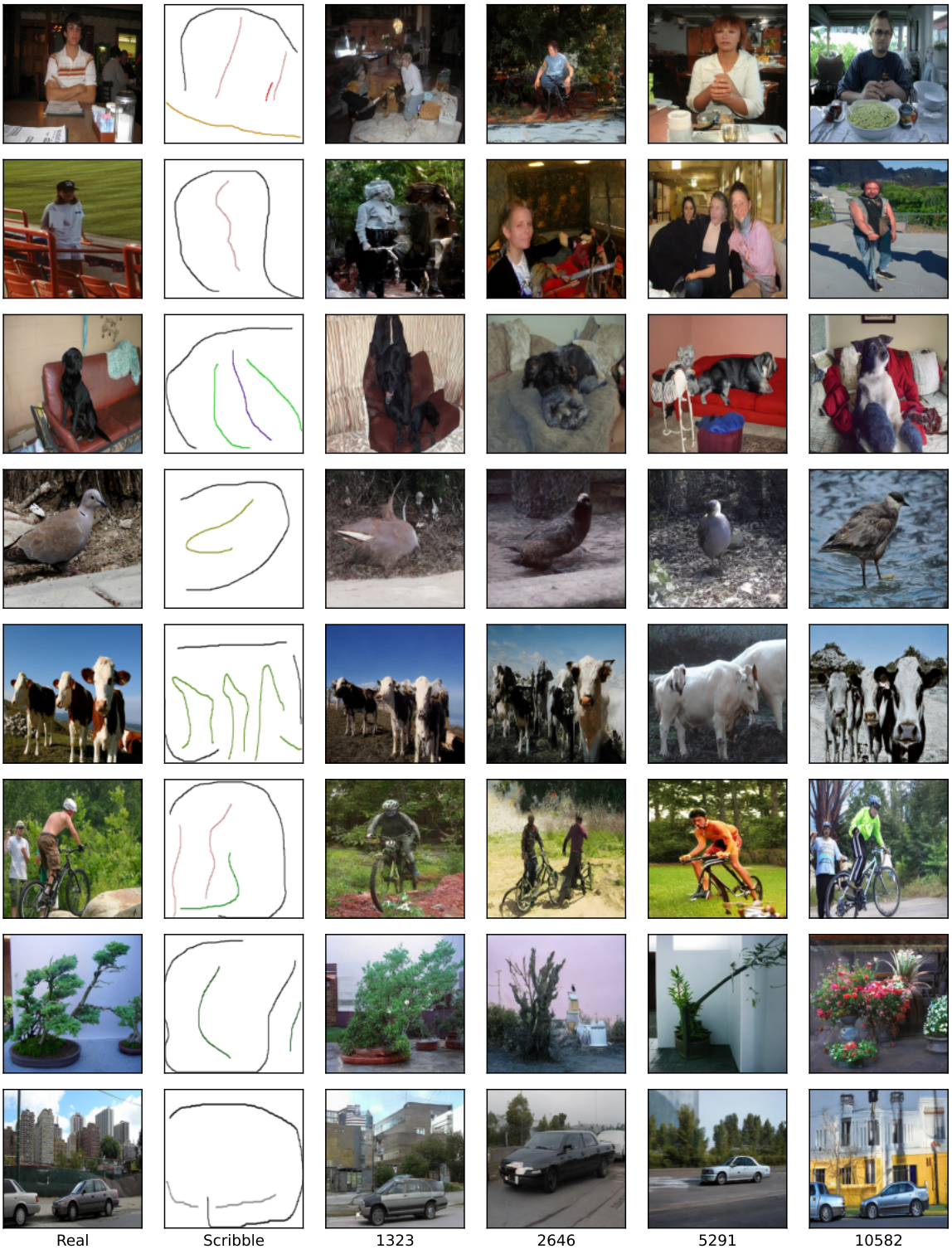

Figure 12. Synthetic training images sampled from a ControlNet model. We vary the number of images on which the ControlNet model is trained to see the impact of the number of training images on synthesis. All images are sampled using guidance scaled $w = 2.0$ and encode ratio $\lambda = 1.0$.

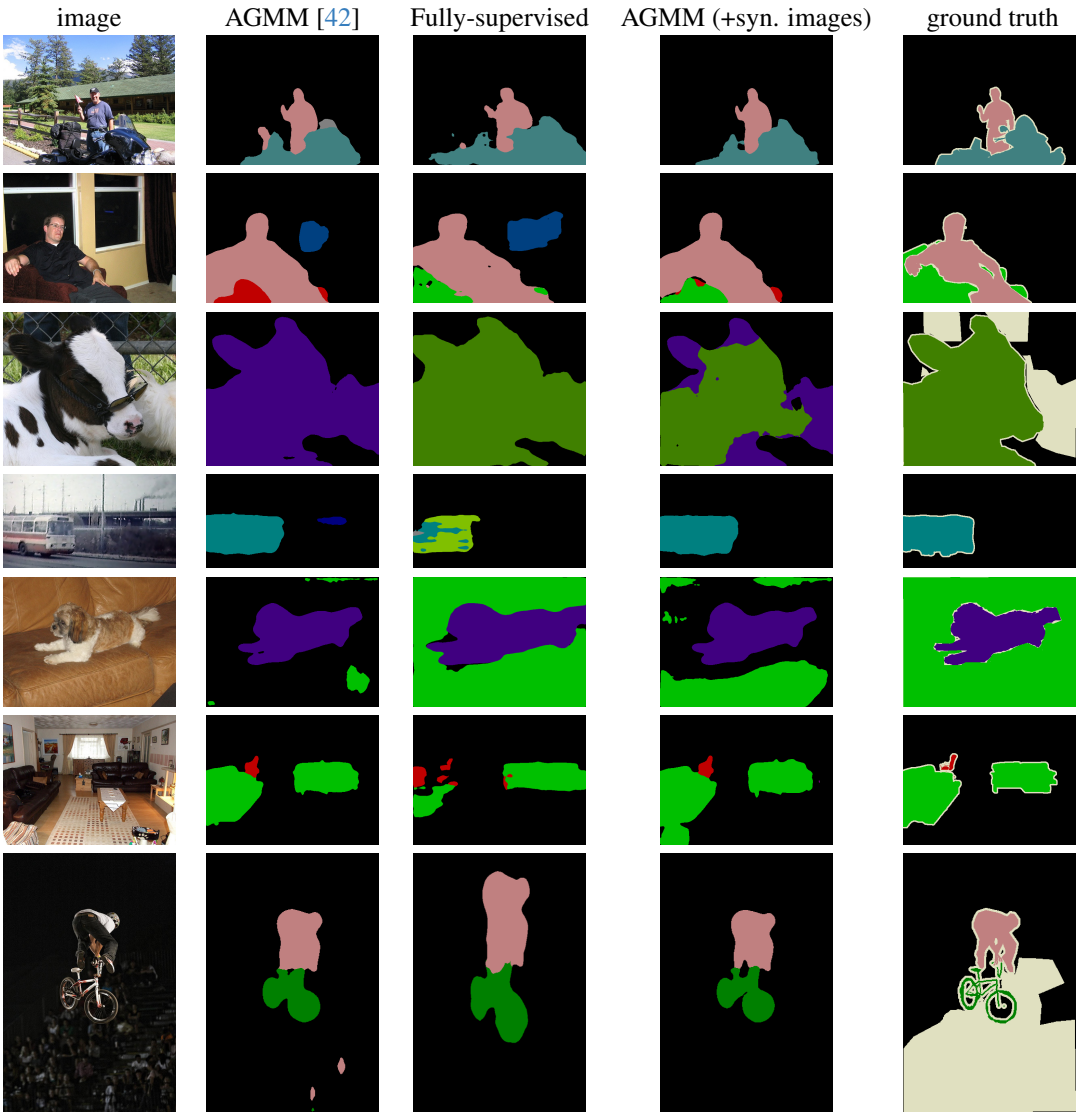

Figure 13. Qualitative results on PASCAL dataset. Our generative data augmentation method improves scribble-supervised semantic segmentation method such as AGMM [42].