Active Learning Inspired ControlNet Guidance for Augmenting Semantic Segmentation Datasets

Hannah Kniesel
Visual Computing Group
Ulm University
hannah.kniesel@uni-ulm.de

Pedro Hermosilla
Computer Vision Lab
TU Vienna
phermosilla@cvl.tuwien.ac.at

Timo Ropinski
Visual Computing Group
Ulm University
timo.ropinski@uni-ulm.de

Abstract

Recent advances in conditional image generation from diffusion models have shown great potential in achieving impressive image quality while preserving the constraints introduced by the user. In particular, ControlNet enables precise alignment between ground truth segmentation masks and the generated image content, allowing the enhancement of training datasets in segmentation tasks. This raises a key question: Can ControlNet additionally be guided to generate the most informative synthetic samples for a specific task? Inspired by active learning, where the most informative real-world samples are selected based on sample difficulty or model uncertainty, we propose the first approach to integrate active learning-based selection metrics into the backward diffusion process for sample generation. Specifically, we explore uncertainty, query by committee, and expected model change, which are commonly used in active learning, and demonstrate their application for guiding the sample generation process through gradient approximation. Our method is training-free, modifying only the backward diffusion process, allowing it to be used on any pretrained ControlNet. Using this process, we show that segmentation models trained with guided synthetic data outperform those trained on non-guided synthetic data. Our work underscores the need for advanced control mechanisms for diffusion-based models, which are not only aligned with image content but additionally downstream task performance, highlighting the true potential of synthetic data generation.

1. Introduction

The rapid advancements in generative models [31–33, 35, 37] have transformed the landscape of data generation, presenting unprecedented opportunities for enhancing training with additional data or even training on synthetic images only [3, 20, 22, 24, 38, 39, 51, 57, 61]. Today, these models can not only be used to generate new data samples but also their corresponding labels [13, 20–22, 24, 30, 52, 57], making them especially valuable for tasks where manual annotation is resource-intensive, such as semantic segmentation. Especially, ControlNet [60] has great potential in this domain, and many techniques for enhancing segmentation datasets build upon it [22, 24, 39, 57]. By being able to generate synthetic training data with labels, manual labeling – traditionally a major bottleneck in computer vision research – is reduced significantly.

Besides the possibility to enhance training datasets without additional labeling cost, generative models enable to guide image content based on predefined principles. This means that unlike real-world images, where we are limited to selecting training samples from a fixed set, this allows us to guide the image generation process toward desired samples when working with synthetic images. In this context, a fundamental question arises: Can we guide the generation process to produce samples that are inherently more valuable for training?

Targeted image generation has been explored in prior work to improve alignment with text prompts or enhance image quality [6, 8, 12, 13, 13, 16, 21, 23, 35, 47]. However, to the best of our knowledge, none of these approaches directly aim to guide sample generation toward creating more useful samples for downstream model performance.

In contrast, various methods [1, 15, 25, 27, 46, 58, 62] have used subsampling or data selection to improve downstream training efficiency and performance. Subsampling techniques have been studied for both real-world and synthetic data based on a range of criteria. More specifically, subsampling a large data pool to extract the most valuable samples directly intersects with active learning [41], a strategy that prioritizes the most informative samples for annotation, improving model performance while minimizing labeling effort. Active learning has proven effective for real-world images as well as synthetic data by reducing annotation costs and computational overhead as well as improving model performance [18, 29, 34, 40, 54, 55, 62].

Building on these approaches, we propose a new research direction: As generative models improve, the need for additional guidance toward realism diminishes. Instead, we focus on guiding the generation process toward the most informative samples for downstream task performance. This shifts the focus from post-generation filtering to proactive, targeted generation of valuable synthetic data.

Within this paper, we investigate whether active learning principles – such as sample difficulty and model uncertainty – can be leveraged to guide the generation of synthetic datasets for semantic segmentation. Specifically, we examine three common active learning query strategies: First, we assess *uncertainty* [26] by using entropy [45]. Second, we approximate *query by committee* [44] by leveraging Bayesian approximation with Monte Carlo dropout layers [17]. Lastly, we apply *expected model change* [43] by computing the cross-entropy loss. Our findings demonstrate that incorporating active learning strategies into the generation process can lead to more informative samples, ultimately improving the downstream performance of segmentation models.

Our approach is innovative in that it remains training-free and directly guides the diffusion process of generative models during inference, targeting data generation in a more strategic manner. Specifically, we contribute the following:

- We introduce active learning inspired ControlNet guidance of the backward diffusion process, to obtain synthetic images which are more valuable for downstream task performance.
- We reveal that entropy is able to outperform other active learning inspired metrics for the proposed ControlNet guidance.
- Our evaluation shows that training on guided synthetic data is able to outperform training on non-guided synthetic data.

In summary, this work presents a novel approach for generating synthetic data optimized for semantic segmentation, leveraging the unique properties of synthetic data and generative models to strategically enhance dataset quality for downstream model training.

2. Related Work

Within this section we highlight relevant works in the field of diffusion models, diffusion models for segmentation, guidance of diffusion models and active learning, as these domains lay the foundation for our ControlNet Guidance.

2.1. Conditional Image Generation

Diffusion Models. Diffusion models, such as *Stable Diffusion* [35], *DALL-E* [33], and *Imagen* [37], have gained attention for generating high-quality samples by learning to reverse the process of adding noise to images. These models also enable conditional generation, often through text prompts. Latent diffusion models [35] further reduce computational complexity by operating in latent space using a VAE. Diffusion models are applied in synthetic image generation, dataset augmentation, and improving adversarial robustness. For example, Azizi et al.[3] used Imagen [37] for augmenting datasets, Zhou et al.[61] showed how Stable Diffusion-based augmentation boosts classification accuracy and Sarıyıldız et al. [38] demonstrated that Stable Diffusion could produce synthetic training data with generalization capabilities comparable to models trained on real data. Recent studies also highlight their potential in adversarial defense [51].

Diffusion Models for Segmentation. Diffusion models have shown promise in segmentation tasks. For instance, *DiffuMask*[52] uses text-guided cross-attention for pixel-wise mask generation, achieving state-of-the-art results in open-vocabulary segmentation. Similarly, *Dataset Diffusion* improves segmentation quality by leveraging self- and cross-attention maps [30].

ControlNet [60] enables conditioning Stable Diffusion on structured data like edges and depth, providing more control over the generation process. This has been applied to data augmentation for segmentation, as in *SegGen* [57], which generates images conditioned on segmentation masks, and *ScribbleGen* [39], which improves segmentation with scribble-based guidance. Similarily, *DGInStyle* [22] enhances data diversity for domain-generalizable segmentation leveraging ControlNet.

Despite the additional conditioning capabilities of ControlNet, Kupyn and Rupprecht [24] found that diffusion models struggle to generate complex scenes with multiple foreground and background objects. To address this limitation and ensure precise alignment between generated images and their corresponding ground truth masks, they propose a structured approach: first, annotations are decomposed into per-object binary masks. Then, an inpainting ControlNet, conditioned on edge and (predicted) depth maps, is used to redraw each instance individually. Finally, the instances are recombined into a coherent scene using depth-based alpha blending. This method significantly improves model training across various tasks, including object detection, semantic segmentation, and instance segmentation, yielding strong results.

While these approaches show promise, they don't explicitly optimize segmentation model performance during data generation, which motivates our approach to directly guide the generation process with ControlNet Guidance for better segmentation outcomes.

Diffusion Model Guidance. Diffusion model guidance has been studied in a large variety of tasks. One of the most prominent guidance methods for diffusion models is

classifier-guided diffusion [13]. In this approach, an auxiliary classifier is trained to assess whether an intermediate diffusion step aligns with the desired attributes, allowing it to guide the generation process for more directed outputs. However, classifier guidance requires the training of an additional classifier, adding computational overhead. Instead, Ho et al. [21] proposed *classifier-free guidance*: This approach eliminates the need for the extra classifier network, by directing the image generation process using a conditional input, such as a text prompt, thereby improving the relevance and coherence of generated images.

In another line of work, guidance has been shown crucial in achieving better alignment between image generation and text prompts. *Training Free Layout Guidance* [8] addresses layout issues by using cross-attention and bounding box prompts during inference to improve layout alignment. Similarly, Singh and Zheng [47] detect and prioritize misaligned parts of generated images within a modified reverse diffusion process, they are able to text-to-image alignment. Feng et al. [16] improve compositional synthesis by using scene graphs to capture spatial and relational aspects in prompts. Additionally, Chefer et al. [6] introduce *Generative Semantic Nursing (GSN)* to amplify activations associated with prompt tokens, improving image synthesis accuracy.

Further, guided diffusion is also employed in generating adversarial images for classification tasks, producing samples with specific properties. Methods like *AdvDiff* [12] andChen et al. [7] generate imperceptible, transferable adversarial attacks using guidance. *AdvDiffuser* [9] synthesizes natural adversarial examples, while Chen et al. [10] propose a content-based attack for precise control over adversarial generation. Guided diffusion is also employed for adversarial purification, helping remove perturbations from images affected by attacks. For instance, Lin et al. [28] propose a purification method using diffusion models to counter adversarial effects, improving model robustness.

These works underscore the potential of guided diffusion to refine prompt alignment, produce controlled adversarial samples, and purify images impacted by adversarial perturbations. Motivated by this potential, within this paper we introduce the first guidance, which directly targets downstream performance of semantic segmentation models, trained on synthetic data.

2.2. Active Learning

Active learning is a machine learning technique designed to enhance model performance while reducing labeling costs by strategically selecting the most informative samples from an unlabeled dataset. This technique focuses on samples that provide the greatest improvement to the model, making it especially valuable when labeling is costly or time-consuming. Active learning has been successfully applied across various domains, from traditional machine learning [42, 43, 49] to large language models [2, 5, 50] and deep learning for computer vision [53, 56, 59].

Key query strategies in active learning include *uncertainty*[26], *query by committee*[44], and *expected model change* [43]. In the uncertainty-based strategy, entropy [45] is often used to measure the model's uncertainty over predicted classes. A high entropy indicates that the model is unsure about its prediction, making the sample more informative. In the query by committee approach, multiple models make predictions on the same input, and if there is significant disagreement among the models, the sample is considered informative and selected for further annotation and model training. However, this approach can be computationally expensive due to the need for a large committee. To mitigate this, Monte Carlo dropout [17] can be used as a Bayesian approximation, effectively simulating the behavior of a committee with a smaller computational overhead. The expected model change strategy focuses on selecting samples that are expected to cause the largest change in the model's parameters. This is typically done by selecting those with the highest computed error, which indicates the greatest potential for improvement.

3. Background

In the following, we describe the basic principles of diffusion models and ControlNet, as these methods form the foundation of our proposed active learning-inspired ControlNet guidance.

3.1. Diffusion Models

Diffusion models are a class of generative models that produce samples by gradually transforming Gaussian noise into a structured image through an iterative denoising process. In the standard latent diffusion framework, a model generates samples by progressively refining a noise sample $\mathbf{x}_T \sim \mathcal{N}(0, 1)$ over T denoising steps until it produces a final, clean image $\mathbf{x}_0$.

At each timestep $t \in [0, T]$, the latent representation $\mathbf{x}_t$ consists of a mixture of an underlying clean image estimate $\hat{\mathbf{x}}_0$ and a noise component $\boldsymbol{\epsilon}_t$. A trained denoising model, typically a U-Net, predicts the noise $\boldsymbol{\epsilon}_t$, which can then be subtracted from $\mathbf{x}_t$ to estimate the original clean image. The denoising update step in a standard diffusion model is given by:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \boldsymbol{\epsilon}_t \right), \tag{1}$$

where α_t represents the noise schedule, which controls the level of noise at each timestep. The full denoising process is performed sequentially over multiple steps, gradually removing noise and refining the image structure.

While diffusion models have demonstrated strong generative capabilities, their standard formulation lacks direct control over the generated content, which can be problematic for tasks requiring precise structure or guidance. This limitation is addressed by ControlNet, which introduces additional conditioning mechanisms to steer the generation process.

3.2. ControlNet

ControlNet [60] extends diffusion models by enabling explicit structural conditioning through auxiliary inputs such as edge maps, segmentation masks, depth maps, or other image features. During training, ControlNet learns a mapping from these control inputs to the desired image output while leveraging the pre-trained weights of a standard diffusion model, such as Stable Diffusion [35]. By incorporating structural priors, it enhances the quality and alignment of generated images, making it particularly valuable for data augmentation in segmentation tasks.

Despite its effectiveness, Kupyn and Rupprecht [24] found that ControlNet struggles with generating complex scenes containing multiple foreground and background objects. To address this, they proposed training and applying ControlNet in an inpainting fashion, redrawing individual instances separately to improve generation fidelity.

4. Method

The core of our proposed method integrates an active learning-inspired guidance mechanism into ControlNet, refining its image generation to produce the most informative samples for downstream model training. By leveraging a pretrained ControlNet, we ensure high-fidelity image generation while prioritizing sample informativeness. Our training-free approach seamlessly incorporates guidance into the backward diffusion process, modifying only three lines of code—highlighting its simplicity and elegance.

4.1. Overview

Figure 1 illustrates our approach, which follows an iterative process of model training, guided data generation, and retraining, akin to nowadays active learning frameworks. We begin by training a segmentation model on the initial dataset. We then adapt the ControlNet introduced in [24] with our proposed active learning inspired ControlNet guidance. Thus we incorporate the segmentation model's predictions to guide the generation process, targeting samples that are expected to be most beneficial for improving model performance. This approach aligns with active learning principles, where the most informative or uncertain samples are selected for retraining [41]. Finally, the generated data is used to augment the original dataset, and the segmentation model is retrained, progressively improving its accuracy and robustness.

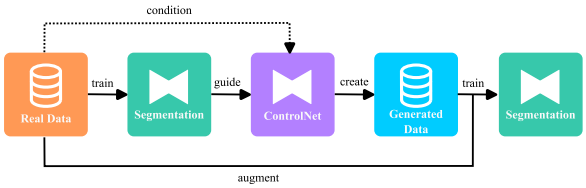


Figure 1. Our proposed iterative data generation and model refinement pipeline introducing active learning inspired ControlNet guidance: A segmentation model trained on real data guides ControlNet to generate informative, real-data-aligned samples, which are then added to the training set for model retraining. This active learning-inspired process refines the model through data generation.

4.2. ControlNet Guidance

Similar to classifier-free guidance [21], we modify the latent representation x_t at each diffusion denoising step $t \in [0, T]$ to drive it towards a more desirable representation $\hat{x}_t$. This approach enforces specific characteristics in the generated data. In the following we will refer to this technique as *latent guidance*. In this paper, we present, for the first time, how to apply latent guidance to ensure that ControlNet-generated images conform to constraints that improve downstream task performance.

Guidance via Gradient Optimization. During latent guidance, we modify the current latent representation $\mathbf{x}_t$ to steer the final image $\mathbf{x}_0$ towards a predefined objective. This is done by introducing an additional gradient-based update:

$$\hat{\mathbf{x}}_t = \mathbf{x}_t - \eta_t \nabla_{\mathbf{x}_t} \mathcal{L}(\hat{\mathbf{x}}_0),$$

where $\mathcal{L}(\hat{\mathbf{x}}_0)$ represents a task-specific loss function computed on the estimated clean image $\hat{\mathbf{x}}_0$. The step size η_t controls the strength of the guidance at each denoising diffusion step t. We will further refer to this as guidance strength.

Guidance Loss. To guide the generative process, we incorporate feedback from a pre-trained segmentation model, g_ϕ, which is trained on real data and provides predictions for synthetic samples. This model allows us to evaluate synthetic data by measuring sample difficulty as well as uncertainty inspired by active learning practice.

To do so, we investigate three commonly used active learning query strategies for guiding the image generation. We investigate query by committee, expected model change and uncertainty [41].

For computing the *query by committee*, we use Monte Carlo dropout as Bayesian approximation, simulating the behavior of a committee. We compute the disagreement of the committee by the variance of model predictions across multiple stochastic forward passes with activated dropout layers. This variance is given by:

$$\text{MCD} = \frac{1}{N} \sum_{n=1}^{N} \left(g_\phi^{(n)}(\hat{\mathbf{x}}_0) - \bar{g}_\phi(\hat{\mathbf{x}}_0) \right)^2$$

where $g_\phi^{(n)}(\hat{\mathbf{x}}_0)$ is the prediction in the n-th forward pass, and $\bar{g}_\phi(\hat{\mathbf{x}}_0)$ is the mean prediction over N passes. A high variance indicates high disagreement of the committee usually highlighting more informative samples.

To guide based on the *expected model change* we employ the Cross-Entropy (CE) loss, measuring the discrepancy between the model's predicted segmentation mask $g_\phi(\hat{\mathbf{x}}_0)$ for the input image $\hat{\mathbf{x}}_0$ and the ground truth mask $\mathbf{y}$. Samples which cause a high CE loss are expected to cause the largest change in the model's parameters. It is defined as:

$$\text{CE} = -\sum_c \mathbf{y}_c \log g_\phi(\hat{\mathbf{x}}_0)_c$$

where c represents the class index. By leveraging this metric, we aim to maximize sample difficulty by generating images where the model prediction deviates significantly from the ground truth.

Finally, for guiding the image generation using *uncertainty*, we apply entropy, which quantifies the uncertainty over predicted class probabilities. Thus, we guide the generation towards high entropy, making it difficult for the model to confidently assign a single class label. The entropy is defined as:

$$\text{Entropy} = -\sum_c g_\phi(\hat{\mathbf{x}}_0)_c \log g_\phi(\hat{\mathbf{x}}_0)_c$$

where high entropy corresponds to greater uncertainty in the model's predictions.

In the following we will refer to the different guidance loss functions as *MCD*, *CE* and *Entropy* respectively.

Gradient Approximation for Guidance. The introduced guidance functions all depend on the segmentation model predictions on the clean image x_0. This clean image is not available for each denoising step. Instead, we estimate the clean image $\hat{x_0}$ within each denoising step based on the current latent x_t, following Song et al. [48].

$$\hat{\mathbf{x}}_0 \approx \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \boldsymbol{\epsilon}_t \right),$$

We refer to this as *single step denoising* shown in Figure 2.

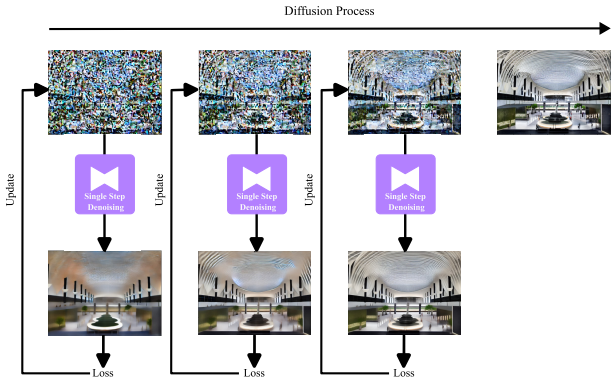


Figure 2. Visualization of latent guidance with single step denoising. For each denoising step, we approximate the clean image $\hat{x_0}$ by single step denoising, such that we are able to apply the loss function to update the current latent x_t.

Algorithm. Our latent guidance approach introduces minimal computational overhead to the standard backward diffusion process. In the following latent guidance pseudocode, we highlight the three additional computations required to apply our active learning-inspired ControlNet guidance in green. These three steps are the only modifications needed compared to the standard backward diffusion process.

Algorithm 1 Latent Guidance in Diffusion Models

Input: Trained UNet model, initial noise sample $\mathbf{x}_T \sim \mathcal{N}(0, 1)$, guidance strength η,
for $t = T, \ldots, 1$ **do**
  $\boldsymbol{\epsilon}_t = \text{UNet}(\mathbf{x}_t, t)$ // Predict noise
  $\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \boldsymbol{\epsilon}_t \right)$ // Estimate clean image
  $\mathbf{g}_t = \nabla_{\mathbf{x}_t} \mathcal{L}(\hat{\mathbf{x}}_0)$ // Compute guidance gradient
  $\hat{\mathbf{x}}_t = \mathbf{x}_t - \eta_t \cdot \mathbf{g}_t$ // Apply guidance
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\hat{\mathbf{x}}_t - \boldsymbol{\epsilon}_t \right)$ // Perform denoising step
end for
Output: Final generated image $\mathbf{x}_0$

5. Experiments

5.1. Experimental Setup

Datasets. We evaluate our approach on well-established benchmark datasets for semantic segmentation, specifically COCO-Stuff10k [4] (COCO10k), which involves predicting 171 different classes, and PascalVOC2012 [14] (VOC2012), with 20 distinct classes. For the ablation studies, we create a reduced version of COCO10k by randomly subsampling it to 2,500 training images, referred to as COCO2.5k.

Evaluation Metric. To assess the impact of our guided image generation, we train a downstream segmentation model using the synthetic data as augmentations. We evaluate segmentation performance using two standard metrics: the mean Intersection over Union (mIoU) and the mean accuracy (mAcc). The mIoU measures the average overlap between predicted and ground truth regions, while mAcc calculates the average accuracy across all classes, providing a balanced view of model performance in semantic segmentation.

Data Generation. As stated in section 2.1 there are many approaches which have shown the usefulness of ControlNet for synthetically augmenting segmentation datasets. In particular, Kupyn and Rupprecht [24] achieved impressive results by addressing ControlNet's limitations in generating complex scenes with multiple objects. Hence, in our work, we follow their methodology and extend it to the proposed active learning inspired guidance. We employ their published ControlNet model with a Stable Diffusion backbone, conditioned on edge maps and (predicted) depth maps. For data generation, we follow their configuration and use $T = 40$ denoising steps. We extract per-object binary masks using connected components [36] from the annotations. Unless stated otherwise, we select single instances for augmentation based on their size, prioritizing the largest objects in the image. Using this procedure, we generate one augmented image per real image, and train the segmentation model with an augmentation probability of $p = 0.5$. If not stated otherwise, within our experiments, we explored guidance scales $\in [7, 10, 20, 30]$ for all datasets and report the best performing. We show the influence of the guidance strength and ablate different guidance schedulers in the supplementary material (see section 7 and Figure 9). For more details please see our codebase (released upon acceptance).

Segmentation Model. We use SegNext-L [19] as our segmentation model, implemented within the MMSegmentation [11] framework. Within our method (Figure 1) we always train the segmentation model from scratch to allow a fair comparison to other approaches. For further training details see our supplementary material or our codebase (released upon acceptance).

5.2. Main Results

In our main results, we compare our method of generating synthetic augmentations by simultaneously guiding the image generation process to produce more useful training samples with the corresponding non-guided approach introduced by Kupyn and Rupprecht [24] at ECCV24. Additionally, we report results for the baseline model, which does not use any additional synthetic data. The methods are evaluated on two commonly used benchmark datasets: VOC2012 and COCO10k.

Although Kupyn and Rupprecht [24] demonstrate promising improvements over the baseline, our method achieves twice the performance gain across all datasets, establishing a new state-of-the-art. The improvements on COCO10k are particularly substantial, showcasing the effectiveness of our approach. On VOC2012, where overall performance is already exceptionally high, further gains are inherently more challenging due to a possible performance ceiling. Nonetheless, our method still manages to achieve measurable improvements, highlighting its robustness even in near-saturated benchmarks.

5.3. Ablations

Guidance Loss. Within this ablation, we investigate the suitability of different guidance metrics as defined in section 4. Despite the overall promising qualitative results in Figure 3 we found one metric was outperforming the others: Table 2a shows the advantage of entropy. We attribute the advantage of entropy to two key limitations of the other metrics in the context of guidance. First, the randomness inherent in the MCD metric makes gradient-based optimization challenging, leading to noisy results, as illustrated in Figure 4. This figure visualizes the model's uncertainty on the generated data (y-axis) across different guidance strengths (x-axis) for each metric. Ideally, uncertainty should increase with stronger guidance, which holds true for both CE and entropy. However, MCD exhibits significant noise, making it unreliable for gradient-based optimization. This instability stems from the stochastic nature of MCD, which introduces variance into the optimization process, reducing its effectiveness.

Second, the CE loss formulation leads to an ill-posed optimization problem. While it encourages the generation of challenging samples—i.e., images where the model confidently predicts incorrect classes—it can also produce samples containing entirely incorrect classes, as demonstrated in Figure 5. This unintended behavior reduces its effectiveness as a reliable guidance signal.

Hence, in all following experiments, we will leverage entropy guidance. For more qualitative evaluations please see our supplementary material at Figure 7 and Figure 8.

Object Selection. As mentioned above, Kupyn and Rupprecht [24] propose redrawing each object individually and recombining them into the final image, ensuring perfect alignment with the existing ground truth mask. While highly effective, this approach is computationally expensive, especially for images containing many objects. To mitigate this, we explore alternative selection strategies that prioritize the most impactful objects for augmentation, aiming to maximize improvements in downstream model train-

Dataset	Method	Syn	Guided	mIoU	mAcc
COCO10k	Baseline	×	×	43.59	55.57
	ECCV24 [24]	✓	×	44.05 (+0.46)	55.88 (+0.31)
	Ours	✓	✓	**45.44** (+1.85)	**56.91** (+1.34)
VOC2012	Baseline	×	×	91.42	95.07
	ECCV24 [24]	✓	×	91.69 (+0.27)	95.09 (+0.02)
	Ours	✓	✓	**91.91** (+0.49)	**95.19** (+0.12)

Table 1. By incorporating guided synthetic data augmentation, our approach more than doubles the performance gain of Kupyn and Rupprecht [24] on COCO10k, achieving an impressive +1.85 mIoU improvement over the baseline. Even on VOC2012, where performance is already near saturation, our method continues to push the boundaries, demonstrating measurable gains, again doubling the performance gain of Kupyn and Rupprecht [24]. These results highlight the effectiveness and generalizability of our approach in enhancing downstream model performance.

	MCD	CE	**Entropy**
mIoU	39.97	39.78	**40.19**
mAcc	50.9	50.62	**51.10**

(a)

	Most Certain	**Largest**
mIoU	39.45	**40.19**
mAcc	50.93	**51.10**

(b)

Object Count	1	**3**
mIoU	40.19	**40.4**
mAcc	51.10	**51.55**

(c)

Table 2. Ablations of the active learning inspired ControlNet guidance by comparing downstream model performance when training with synthetically augmented data. (a) Comparison of different guidance loss functions (MCD, CE, Entropy), where entropy shows the most promising results. (b) Evaluation of augmenting the largest instance with guidance compared to augmenting the most certain instance. Augmenting the largest instance is a natural choice as it usually contributes more to the training loss. (c) Comparison of augmenting different numbers of objects with our method. Consistent with the ablation on object selection, augmenting larger regions of the image enhances performance.

ing. We focus on selection criteria based on object size and model uncertainty. Larger objects typically contribute more to the training loss, making them ideal for augmentation, so we prioritize selecting the largest objects. Additionally, inspired by active learning, we consider augmenting instances where the model is most certain. These objects only contribute little to the loss, and hence are expected to bring the greatest improvement when augmented, further enhancing the training process.

As shown in Table 2b, selecting the largest instance remains more beneficial than selecting the most certain instance. This is likely because the most certain instances are often smaller, meaning their guided uncertainty would need to be significantly higher for them to have an equivalent impact than larger objects.

Number of Objects. The ablation study on instance selection suggests that augmenting larger objects leads to better model performance. To further explore this, we examine the effect of varying the number of augmented objects, which consequentially leads to larger augmented regions. As shown in Table 2c, our findings align with the previous experiment: augmenting larger regions of the image, hence augmenting more instances, improves performance.

6. Conclusion

In this work, we presented a novel paradigm for guiding the generation of synthetic training data introducing active learning inspired ControlNet guidance. By incorporating ControlNet guidance based on active learning criteria, we demonstrated that the value of synthetic data for downstream tasks can be substantially increased. Our results show that actively steering the diffusion process toward generating informative samples leads to superior segmentation model performance compared to unguided data generation. This shifts the conventional approach from post hoc sample selection to proactive, task-aware data synthesis, fundamentally redefining how synthetic datasets are constructed.

Our results highlight that entropy-based guidance outperforms other criteria, emphasizing the importance of uncertainty-based selection in generative modeling. However, our method introduces computational overhead due to segmentation model training and guidance, motivating research into efficient surrogates. Additionally, the strength of the guidance is a crucial hyperparameter—if too weak, the effect is negligible; if too strong, the generated images may deviate from the original distribution. While this could impact immediate model performance, it may also enhance generalization, which remains an avenue for future work.

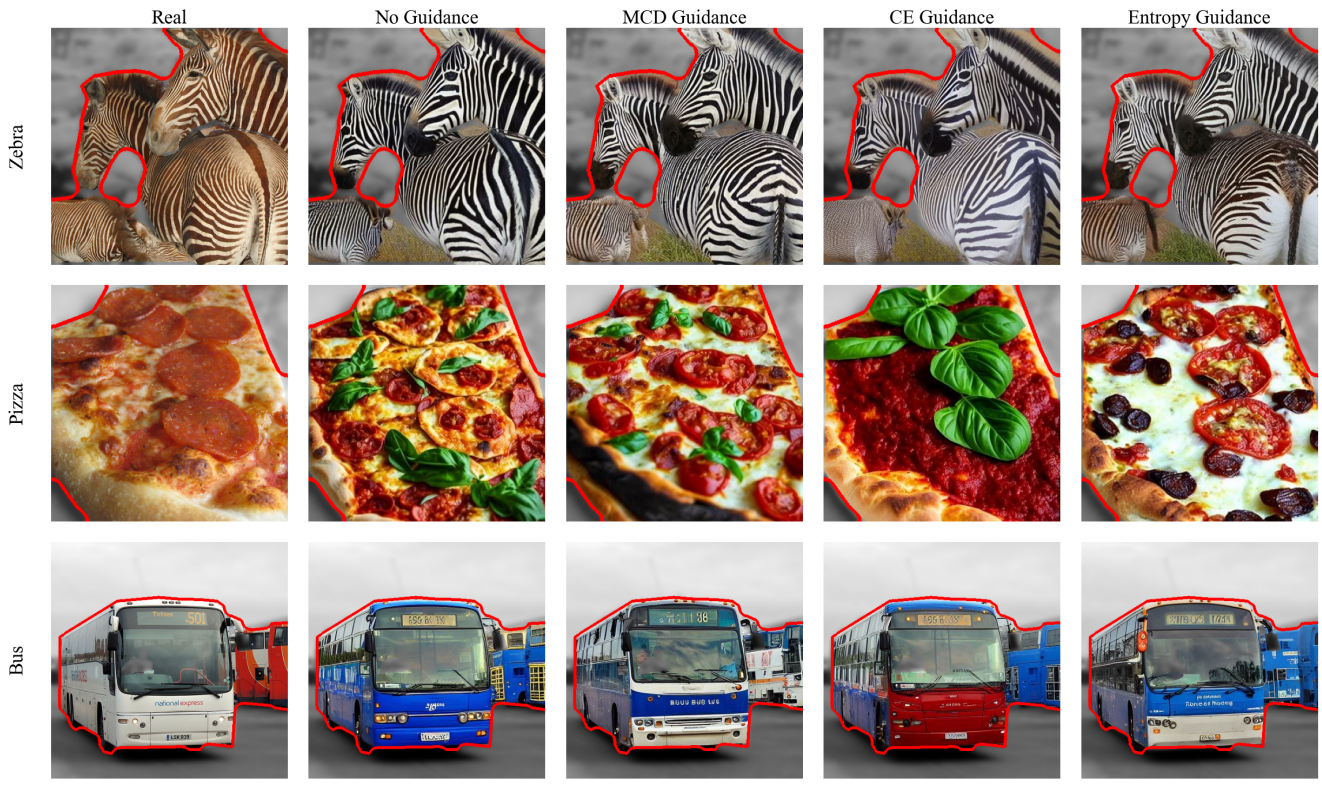


Figure 3. Qualitative comparison of different loss metrics during guidance. We visualize the real images (top row) as well as synthetically augmented images following [24] (second row) next to our proposed guidance in the following rows. Red borders outline the synthetically augmented object. The images share high visual quality.

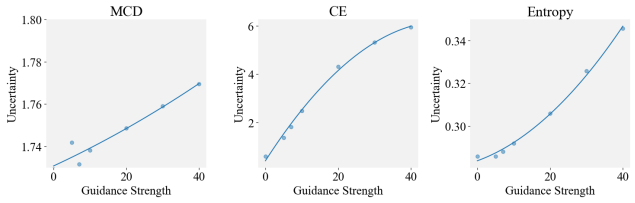


Figure 4. The plots visualize the uncertainty of generated objects when applying guided backward diffusion using different metrics at varying guidance strengths. Uncertainty is measured according to the respective metric. As guidance strength increases, we expect uncertainty to rise accordingly. However, MCD introduces noise, making it unreliable for gradient-based optimization. This instability arises from its stochastic nature, which injects variance into the optimization process, ultimately reducing its effectiveness.

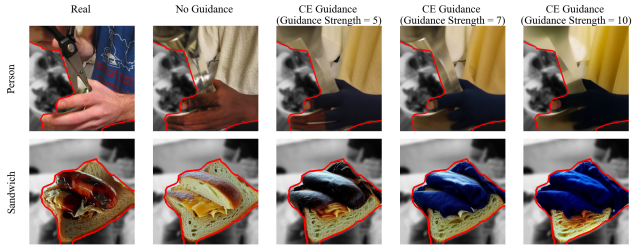


Figure 5. Visualization of failure cases of CE guidance. Guiding based on CE leads to an ill-posed optimization problem. It can encourage the prediction of images which are being misclassified by the model, but it might as well predict samples containing incorrect classes.

Beyond these considerations, our work lays the foundation for a broader research direction. Future studies could, again inspired by active learning, investigate iterative, cyclic approaches where the segmentation model continuously refines itself through newly generated data, akin to curriculum learning. This could unlock further improvements in model robustness and efficiency.

Overall, this work presents a transformative approach to dataset generation, demonstrating that generative models can do more than merely mimic real-world data—they can be harnessed to produce training samples that are intrinsically more valuable. As generative models continue to improve, this perspective will be critical in shaping the future of data-driven deep learning, reducing reliance on expensive manual annotations, and enabling more efficient AI systems.

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022. 1

[2] Nicolás Astorga, Tennison Liu, Nabeel Seedat, and Mihaela van der Schaar. Active learning with llms for partially observed and cost-aware scenarios. *Advances in Neural Information Processing Systems*, 37:20819–20857, 2025. 3

[3] 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. 1, 2

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5

[5] Luckeciano Carvalho Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. Deep bayesian active learning for preference modeling in large language models. *Advances in Neural Information Processing Systems*, 37:118052–118085, 2025. 3

[6] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 1, 3

[7] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[8] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 1, 3

[9] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4562–4572, 2023. 3

[10] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[11] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 6

[12] Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. In *European Conference on Computer Vision*, pages 93–109. Springer, 2025. 1, 3

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 1, 3

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5

[15] Chengxiang Fan, Muzhi Zhu, Hao Chen, Yang Liu, Weijia Wu, Huaqi Zhang, and Chunhua Shen. Divergen: Improving instance segmentation by learning wider data distribution with more diverse generative data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3986–3995, 2024. 1

[16] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1, 3

[17] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. 2, 3

[18] Dibet Garcia, João Carias, Telmo Adão, Rui Jesus, Antonio Cunha, and Luis G Magalhães. Ten years of active learning techniques and object detection: a systematic review. *Applied Sciences*, 13(19):10667, 2023. 1

[19] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. *Advances in neural information processing systems*, 35:1140–1156, 2022. 6

[20] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 1

[21] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 3, 4

[22] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dginstyle: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *European Conference on Computer Vision*, pages 91–109. Springer, 2025. 1, 2

[23] Jae Myung Kim, Jessica Bader, Stephan Alaniz, Cordelia Schmid, and Zeynep Akata. Datadream: Few-shot guided dataset generation. In *European Conference on Computer Vision*, pages 252–268. Springer, 2024. 1

[24] Orest Kupyn and Christian Rupprecht. Dataset enhancement with instance-level augmentations. *arXiv preprint arXiv:2406.08249*, 2024. 1, 2, 4, 6, 7, 8, 3

[25] Andrea Lampis, Eugenio Lomurno, and Matteo Matteucci. Bridging the gap: Enhancing the utility of synthetic data via post-processing techniques. *arXiv preprint arXiv:2305.10118*, 2023. 1

[26] David D Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, pages 13–19. ACM New York, NY, USA, 1995. 2, 3

[27] Wenyan Li, Jonas F Lotz, Chen Qiu, and Desmond Elliott. Data curation for image captioning with text-to-image generative models. 2023. 1

[28] Guang Lin, Zerui Tao, Jianhai Zhang, Toshihisa Tanaka, and Qibin Zhao. Robust diffusion models for adversarial purification. *arXiv preprint arXiv:2403.16067*, 2024. 3

[29] Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 580–588. Springer, 2018. 1

[30] Quang Nguyen, Truong Vu, Anh Tran, and Khoi Nguyen. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[31] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1, 2

[34] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021. 1

[35] 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. 1, 2, 4

[36] Azriel Rosenfeld and John L Pfaltz. Sequential operations in digital picture processing. *Journal of the ACM (JACM)*, 13 (4):471–494, 1966. 6

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1, 2

[38] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8011–8021, 2023. 1, 2

[39] Jacob Schnell, Jieke Wang, Lu Qi, Vincent Tao Hu, and Meng Tang. Scribblegen: Generative data augmentation improves scribble-supervised semantic segmentation. 1, 2

[40] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 1

[41] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 1, 4

[42] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *proceedings of the 2008 conference on empirical methods in natural language processing*, pages 1070–1079, 2008. 3

[43] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20, 2007. 2, 3

[44] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, 1992. 2, 3

[45] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2, 3

[46] Kashun Shum, Yuzhen Huang, Hongjian Zou, Ding Qi, Yixuan Liao, Xiaoxin Chen, Qian Liu, and Junxian He. Predictive data selection: The data that predicts is the data that teaches. *arXiv preprint arXiv:2503.00808*, 2025. 1

[47] Jaskirat Singh and Liang Zheng. Divide, evaluate, and refine: Evaluating and improving text-to-image alignment with iterative vqa feedback. *Advances in Neural Information Processing Systems*, 36:70799–70811, 2023. 1, 3

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[49] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001. 3

[50] Chuqiao Wang, Junru Li, and Ruiming Zhang. A method to enhance structural fairness in large language models with active learning. *Authorea Preprints*, 2024. 3

[51] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pages 36246–36263. PMLR, 2023. 1, 2

[52] 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. 1, 2

[53] Binhui Xie, Longhui Yuan, Shuang Li, Chi Harold Liu, and Xinjing Cheng. Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8068–8078, 2022. 3

[54] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian conference on computer vision*, 2020. 1

[55] Chenhongyi Yang, Lichao Huang, and Elliot J Crowley. Plug and play active learning for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17784–17793, 2024. 1

[56] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 399–407. Springer, 2017. 3

[57] Hanrong Ye, Jason Kuen, Qing Liu, Zhe Lin, Brian Price, and Dan Xu. Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. In *European Conference on Computer Vision*, pages 352–370. Springer, 2025. 1, 2

[58] Jiarong Ye, Yuan Xue, L Rodney Long, Sameer Antani, Zhiyun Xue, Keith C Cheng, and Xiaolei Huang. Synthetic sample selection via reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 53–63. Springer, 2020. 1

[59] Tianning Yuan, Fang Wan, Mengying Fu, Jianzhuang Liu, Songcen Xu, Xiangyang Ji, and Qixiang Ye. Multiple instance active learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5330–5339, 2021. 3

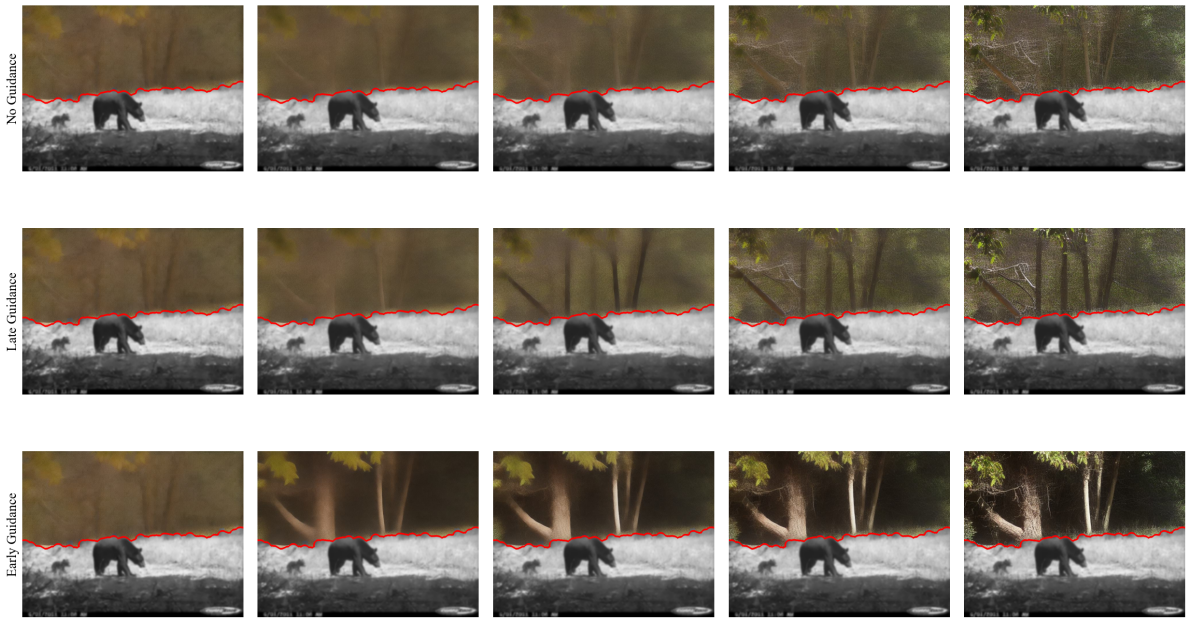
[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1, 2, 4

[61] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Using synthetic data for data augmentation to improve classification accuracy. 2023. 1, 2

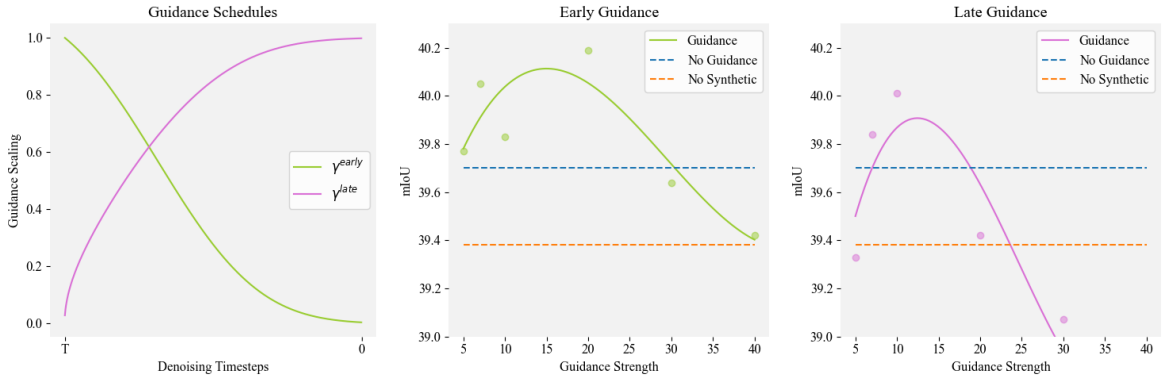
[62] Muzhi Zhu, Chengxiang Fan, Hao Chen, Yang Liu, Weian Mao, Xiaogang Xu, and Chunhua Shen. Generative active learning for long-tailed instance segmentation. In *International Conference on Machine Learning*, 2024. 1

Active Learning Inspired ControlNet Guidance for Augmenting Semantic Segmentation Datasets

Supplementary Material



(a)



(b) (c) (d)

Figure 6. (a) Qualitative assessment of different guidance schedules. The figure shows the approximated $\hat{x_0}$ at different denoising iterations from left to right. The top row depicts unguided backward diffusion, while the middle and bottom rows illustrate guided backward diffusion using γ^{late} and γ^{early}, respectively. Applying γ^{early} can lead to structural modifications in object features, whereas γ^{late} primarily affects texture. (b) Visualization of the early guidance (γ^{early}) and the late guidance (γ^{late}) schedulers to control the guidance strength for each denoising step individually. (c) Downstream model performance when trained on data generated with γ^{early} at different guidance strengths. The orange line represents performance without synthetic augmentation, while the blue line corresponds to performance with synthetic augmentation following [24]. (d) Same as (c), but using γ^{late}. These results suggest that model training benefits more from structural modifications introduced by γ^{early} rather than textural changes. Textural modifications tend to produce unnatural images more quickly, even when imperceptible to the human eye—similar to JPEG compression, which removes high-frequency details due to their limited visibility, or adversarial perturbations, which introduce imperceptible high-frequency changes to manipulate the image—ultimately leading to quicker performance degradation.

7. Guidance Schedule

Guiding the backward diffusion process at different denoising timesteps can yield varying effects on the generated image. We therefore investigate the impact of applying stronger guidance during early iterations compared to later iterations. We base the schedulers on the noise schedule of the forward diffusion process. We define $\gamma^{early}(t) = \alpha_t$ and $\gamma^{late}(t) = \sqrt{1 - \alpha_t}$, where α_t represents the cumulative product of the noise schedule's variance values at time step t. For a visualization of the schedules, see Figure 6b.

As shown in Figure 6a, we observe that γ^{early} significantly alters the structural features of the objects, while γ^{late} primarily affects texture.

In our quantitative experiment, we compare no synthetic data augmentation to non-guided augmentations based on Kupyn et al.'s work [24], and our proposed guided data generation using the entropy loss function. We test different guidance scales with the introduced schedulers and present the results in Figure 6c and Figure 6d. Although the images generated with both schedules appear to have high quality, as shown in Figure 6a, we found that early guidance, γ^{early}, has a more pronounced impact than late guidance, γ^{late} (see Figure 6c and Figure 6d). Furthermore, we observed that the overall guidance strength can be higher for early guidance, which highlights the effectiveness of γ^{early} in generating more natural changes in the image content, preventing overly aggressive, unnatural shifts in the generated images. Notably, when the guidance strength is too high, the generated images deviate from the training/validation data distribution, resulting in performance degradation.

8. Implementation Details

Segmentation Model. For training the segmentation model, we adopt the AdamW optimizer with an initial learning rate of 6×10^{-5}, following a poly learning rate (PolyLR) schedule. We use a batch size of 16. The maximum training iteration is dataset-dependent:

- COCO2.5k: 40,000 iterations
- COCO10k: 80,000 iterations
- VOC2012: 40,000 iterations

Data Generation. For data generation with ControlNet, we utilize a pretrained model from Kupyn and Rupprecht [24] using the Stable Diffusion backbone. Similar to Kupyn and Rupprecht [24] we incorporate $T = 40$ diffusion denoising steps. The guidance is conditioned on predicted depth and HED edge detection, ensuring structure-aware synthesis. Our generation pipeline employs an image size of 768. To balance fidelity and control strength, we vary the guidance scale between 6.0 and 7.5, while the conditioning scale is set at 0.7 and 0.2. As the guidance strength is an important parameter in our active learning inspired ControlNet guidance, we test different guidance strength for the different datasets, where $\eta \in [7, 10, 20, 30]$. We finally used following setup:

- COCO2.5k: $\eta = 20$
- COCO10k: $\eta = 7$
- VOC2012: $\eta = 10$

For further details, please see our code base (released upon acceptance).

9. Visualizations

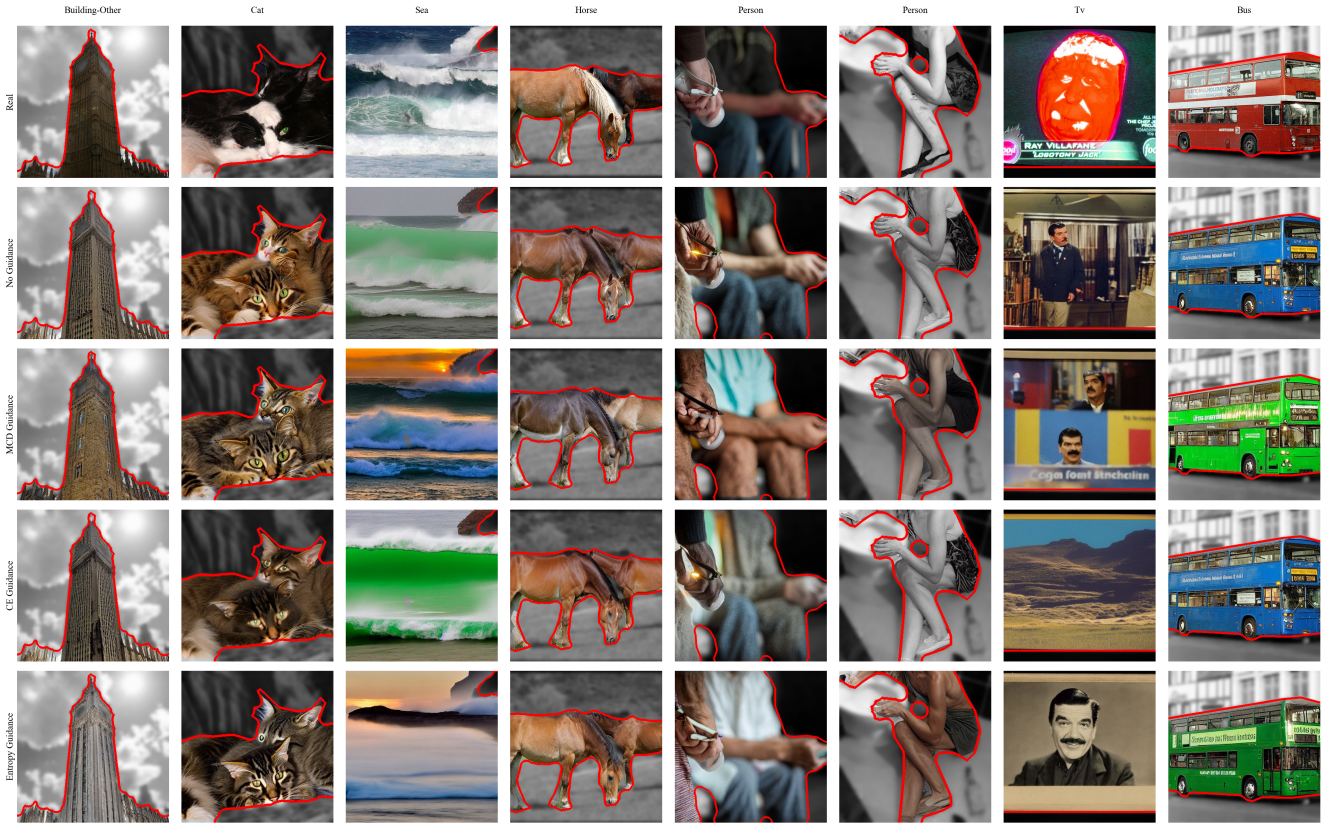


Figure 7. More qualitative results on the comparison of different loss metrics during guidance for multiple classes. We visualize the real images (top row) as well as synthetically augmented images following [24] (second row) next to our proposed guidance in the following rows. Red borders outline the synthetically augmented object. The images share high visual quality.

Figure 8. More qualitative results on the comparison of different loss metrics during guidance for class ”person”. We visualize the real images (top row) as well as synthetically augmented images following [24] (second row) next to our proposed guidance in the following rows. Red borders outline the synthetically augmented object. The images share high visual quality.

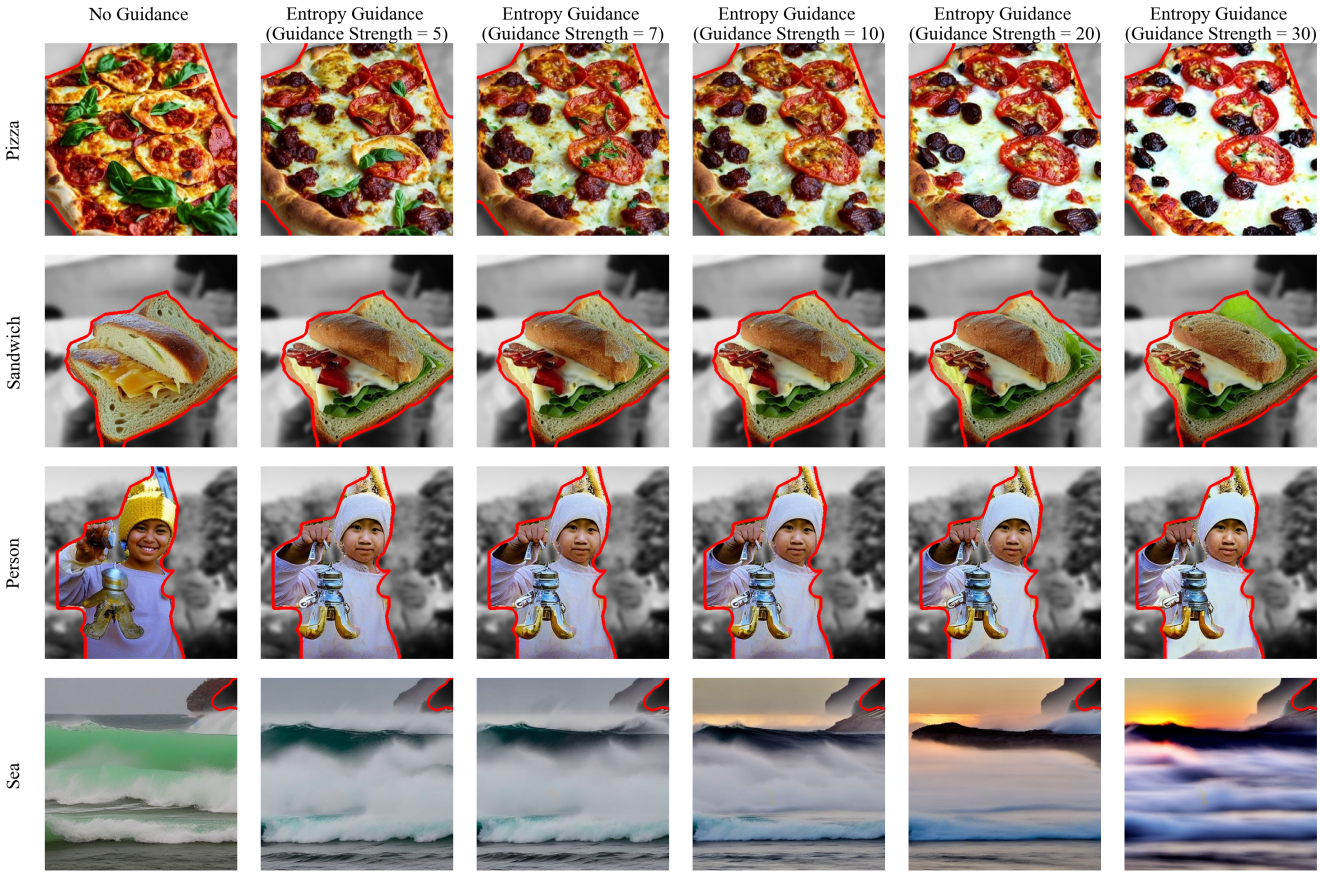


Figure 9. Qualitative results showing the influence of guidance strength. The first column displays the non-guided images, while the subsequent columns show results with our proposed active learning-inspired ControlNet guidance at varying strengths. Red borders highlight the synthetically augmented objects. As the guidance strength increases, more pronounced changes are observed in the images. However, stronger guidance can also lead to unintended alterations in the content, such as the green sandwich bread in the second row for a guidance strength of 30.