R&D: Balancing Reliability and Diversity in Synthetic Data Augmentation for Semantic Segmentation

Quang-Huy Che[1,2], Dinh-Duy Phan[1,2,*], Duc-Khai Lam[1,2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
`huycq@uit.edu.vn, duypd@uit.edu.vn, khaild@uit.edu.vn`

Abstract. Collecting and annotating datasets for pixel-level semantic segmentation tasks are highly labor-intensive. Data augmentation provides a viable solution by enhancing model generalization without additional real-world data collection. Traditional augmentation techniques, such as translation, scaling, and color transformations, create geometric variations but fail to generate new structures. While generative models have been employed to extend semantic information of datasets, they often struggle to maintain consistency between the original and generated images, particularly for pixel-level tasks. In this work, we propose a novel synthetic data augmentation pipeline that integrates controllable diffusion models. Our approach balances diversity and reliability data, effectively bridging the gap between synthetic and real data. We utilize *class-aware prompting* and *visual prior blending* to improve image quality further, ensuring precise alignment with segmentation labels. By evaluating benchmark datasets such as PASCAL VOC and BDD100K, we demonstrate that our method significantly enhances semantic segmentation performance, especially in data-scarce scenarios, while improving model robustness in real-world applications. Our code is available at https://github.com/chequanghuy/Enhanced-Generative-Data-Augmentation-for-Semantic-Segmentation-via-Stronger-Guidance.

Keywords: Synthetic Data Augmentation · Semantic Segmentation · Stable Diffusion.

1 Introduction

Deep learning has transformed the field of computer vision, where model performance depends not only on methodological advancements but also significantly on the quality and quantity of training data. Large-scale datasets, such as SA-1B [1], and Imagenet [2], have played a crucial role in driving progress across various computer vision tasks. However, collecting and annotating these datasets is labor-intensive, especially for complex and privacy-sensitive data. This challenge is particularly notable in semantic segmentation, where each pixel in an

---

[*] Corresponding author

image must be accurately classified. While widely used datasets like PASCAL VOC [3], BDD100K [4] provide a strong foundation for training segmentation models, expanding or creating new datasets of similar scale remains a significant bottleneck. Consequently, data augmentation has emerged as a critical approach to enhancing model generalization without requiring additional real-world data collection. This technique not only increases data diversity but also reduces annotation costs, offering an efficient alternative for addressing challenges in semantic segmentation.

Traditional data augmentation methods such as rotation, scaling, flipping, or pixel-level manipulations (e.g., blurring, adjusting brightness, and contrast) enhance model accuracy by introducing geometric and color variations. However, these transformations do not generate new structural components, perspectives, or textures, thus limiting their ability to expand dataset diversity. More advanced techniques include partial image removal methods (e.g. Random Erasing [5], Cutout), or image mixing techniques (e.g., Mosaic [6], Mixup [7]). However, most of these methods primarily expand the visual representation space without introducing new semantic information, thereby reducing their effectiveness in improving the model's generalization capability.

Unlike previous data augmentation methods [8], generative models are trained directly on the target dataset to produce additional samples. However, since these models learn from the same data domain, the generated samples often lack diversity compared to the original data. Without fine-tuning the target dataset, the synthesized images tend to follow the distribution of the pre-trained model, not the desired distribution for data augmentation. Although generative models [9–11] can generate semantically diverse images, ensuring distributional alignment between the original and generated data remains a challenge. Additionally, semantic segmentation requires that generated samples preserve precise object shapes and structures, unlike classification [12, 13] or object detection tasks [14]. To address these limitations, we propose a synthetic data augmentation pipeline for semantic segmentation based on generative models. In summary, the contributions of our work are as follows:

- We propose a novel synthetic data augment pipeline that integrates two controllable diffusion models to generate synthetic datasets for semantic segmentation. This approach bridges the gap between synthetic datasets and real datasets, ensuring both the diversity and reliability of synthetic images.
- We integrate the proposed pipeline with the *class-aware prompting* method we propose and *visual prior blending* [11]. These combined methods enhance the quality of the generated images by ensuring that all relevant objects are included in the generated images and improving the alignment of the generated images with segmentation labels, thereby ensuring high accuracy and reliability in the synthetic datasets.
- We demonstrate the effectiveness of our approach through extensive experiments on standard benchmarks, including PASCAL VOC [3] and BDD100K [4]. Our method consistently improves semantic segmentation performance, particularly in data-scarce scenarios.

2   Related work

2.1   Image Generation

Image generation is a significant research direction in computer vision and artificial intelligence, especially with the rapid advancement of deep learning models in recent years. Generative Adversarial Networks (GANs) [15], as foundational models in image synthesis, have been widely used to generate high-resolution images. However, GANs often face optimization challenges, making it difficult for the model to fully capture the underlying data distribution. Recently, diffusion models (DMs) have emerged as a more advanced approach to image generation, enabling the model to approximate the data distribution more stably compared to GANs. Stable Diffusion (SD) [16,17] is a variant of diffusion models that leverages the latent space instead of directly processing images in the pixel space. Through the cross-attention mechanism, SD can generate images based on various input modalities such as text, bounding boxes, or semantic maps. One key advancement that enhances controllability in image generation is the integration of SD with ControlNet [18] or T2I-Adapter [19]. These methods allow the model to incorporate additional structured guidances (visual priors), represented as edges, segmentation masks, lineart, and depth maps, improving the consistency of shape and structure in the generated images.

2.2   Image Synthesis for Data Augmentation

Previous studies have utilized Generative Adversarial Networks (GANs) to generate synthetic data for semantic segmentation, focusing primarily on object-centered images. However, these methods face limitations when handling complex image layouts or interactions between multiple objects. With advancements in generative models, data augmentation techniques based on diffusion models have recently emerged. However, most of these methods are tailored for image classification [12, 13] or object detection [14] rather than semantic segmentation, which requires pixel-level precision. Semantic segmentation poses a significant challenge for generative image synthesis due to its strict accuracy requirements. [9, 10] introduced Synthetic Dataset approaches capable of generating synthetic images along with segmentation masks for specified classes. These methods generate synthetic datasets and pseudo-labels from text descriptions, enabling data utilization for pretraining segmentation models. Unlike synthetic dataset approaches, generative-based augmentation uses existing images and masks to create additional training data. Inpainting-based methods [20] modify objects while preserving backgrounds but often limit data diversity. In contrast, Che et al. [11] introduced the Controllable Diffusion Model with strong guidance for image synthesis, demonstrating notable improvements in data augmentation. However, synthetic data generation faces two challenges: (1) mismatches between segmentation masks and synthesized images and (2) domain shifts due to the generative model's training dataset constraints.

In this work, we propose a synthetic data augmentation pipeline based on generative models. Our pipeline integrates advanced techniques to enhance the

robustness of the generated data. Additionally, our method achieves a balance between diversity and data reliability consistency compared to the original dataset, resulting in high-quality synthetic data suitable for training semantic segmentation models.

3   Methods

In this work, we propose a pipeline that integrates two SD models designed for controllable synthetic data generation: the Image-to-Image Controllable Diffusion Model (Sec. 3.2) and the Controllable Inpainting Diffusion Model (Sec. 3.3). This pipeline takes an image and its corresponding segmentation labels as input and generates two synthetic images for each input image. The overall architecture of the proposed pipeline is illustrated in Fig. 1. Given a real dataset $\mathcal{D}_0$, proposed pipeline generates the synthetic dataset $\mathcal{D}_1^{gen} \cup \mathcal{D}_2^{gen}$, where $\mathcal{D}_1^{gen}$ is generated by the Image-to-Image Controllable Diffusion Model, producing a highly diverse dataset by changing both labeled and unlabeled objects as well as the background, while $\mathcal{D}_2^{gen}$ is generated by the Controllable Inpainting Diffusion Model, ensuring data distribution consistency by modifying only the labeled objects while keeping the remaining parts unchanged. The merging of the two datasets $\mathcal{D}_1^{gen}$ and $\mathcal{D}_2^{gen}$ yields a reliable synthetic dataset that simultaneously maximizes diversity and preserves data distribution fidelity. Furthermore, to enhance synthetic image precision, we propose novel methods for textual prompt refinement and visual prior in Section 3.1.

3.1   Robust condition for Diffusion Controllable Models

Preparing text prompt: To generate high-quality synthetic images for semantic segmentation, constructing an effective prompt is crucial in ensuring the presence of all relevant objects in the generated image. A straightforward ap-
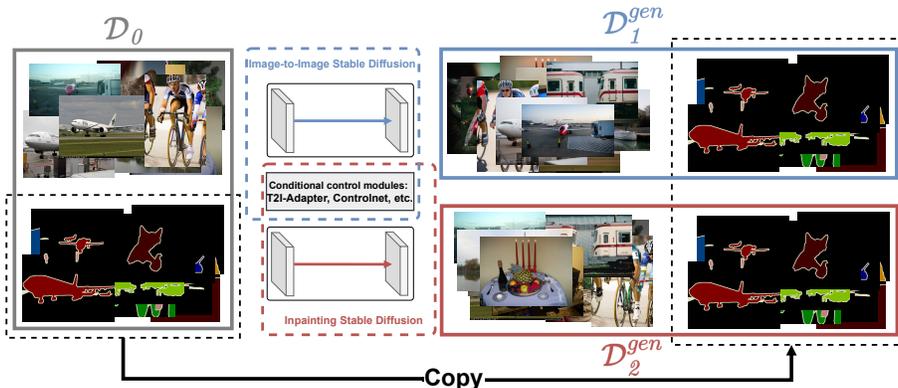

Fig. 1: Our proposed synthetic data augmentation pipeline utilizes the real dataset $\mathcal{D}_0$ to create two synthetic datasets, $\mathcal{D}_1^{gen}$ and $\mathcal{D}_2^{gen}$. The annotations for the synthetic data are directly copied from the labels of the real dataset.

proach is to list the annotated classes explicitly. Given an image $\mathcal{I}_i$ with list of labeled classes $\mathcal{C}_i = [c_1, c_2, \ldots]$, a simple prompt can be formulated as "`A photograph of c`$_1$`, c`$_2$`,...`". While this approach ensures that all objects in the image are mentioned, it lacks contextual information, making it challenging for the generative model to produce a coherent and realistic image. Instead of using simple annotated class lists, another approach is applying image captioning models to generate descriptions for datasets. However, these captions do not guarantee the inclusion of all annotated classes in the image, which may result in generated images that are either incomplete or contain incorrect objects.

To address these challenges, we propose a prompt formulation integrating general contextual information about the image and a list of annotated classes. Unlike previous works [9, 11], which merely concatenate the image caption with the annotated class list—often leading to poor linguistic coherence—we utilize BLIP [21] as an conditional image captioning. Specifically, BLIP generating a more comprehensive description that combines visual context with class labeled list. To enhance the focus on class tokens corresponding to target objects, we propose re-weighting mechanism during class token embedding [22]. By assigning higher weights to class tokens, this approach emphasizes key objects, thereby improving segmentation accuracy. The adjusted class tokens are denoted as "[*class*]++". Our proposed prompt generation method, called *class-aware prompting*, focuses on integrating class-specific information to produce more contextually rich and accurate prompts. Figure 2 illustrates an example of an image alongside its corresponding label, as well as various types of prompts.

Visual priors for controllable model: Controllable generative models are characterized by their ability to generate high-quality images guided by visual priors. Among these, edge-based visual priors are widely used for object representation because they can generalize image structures. However, relying solely on edge information may introduce limitations when target objects are not well-emphasized or edge maps lack sufficient detail. This limitation can lead to generated objects not aligning accurately with the segmentation labels. To address this issue, our proposed pipeline incorporates *visual prior blending* [11], a technique designed to enhance the representation of labeled objects, ensuring that generated content better aligns with segmentation labels. Given V^I as the visual prior derived from the original image and V^S as the visual prior extracted from the segmentation mask, the blended visual prior V^* is formulated as:

$$V^* = \alpha V^I + V^S \tag{1}$$

where $\alpha \in (0, 1)$ is a blending coefficient. Setting $\alpha < 1$ reduces the influence of global image structures while emphasizing the information from labeled objects.

3.2   Image-to-Image Controllable Diffusion Model

Controllable Diffusion Models [18, 19] have demonstrated remarkable capabilities in generating highly diverse synthetic images [11, 14], offering significant

Generated caption: There is a living room with a fireplace and a couch
Simple text prompt: The photograph of chair, sofa
Class-prompt appending: There is a living room with a fireplace and a couch; chair, sofa
Conditional image captioning *(sofa, chair)*: A photograph of chair, sofa, ottoman, and fireplace in a living room
Conditional image captioning *(sofa)*: A photograph of chair, sofa, ottoman, and fireplace in a living room
Conditional image captioning *(chair)*: A photograph of chair, and ottoman in a living room
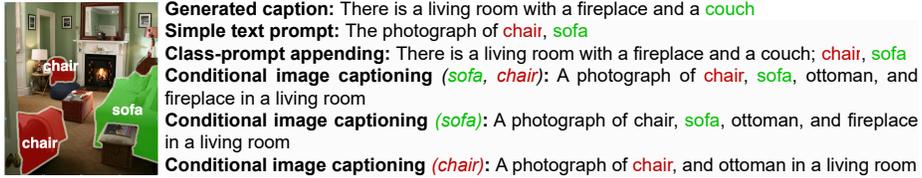
Fig. 2: Some examples of text prompt selection for input images show that simple text prompts are often too simplistic, while generated captions may miss some labeled classes. Class-prompt appending addresses this but can lead to incoherent prompts. In contrast, conditional image captioning creates coherent prompts that accurately describe the image and include all labeled classes.

advantages for data augmentation techniques. The image generation can be process as a function $\mathbf{G}_0 : V \times P \to I^{gen}$, where V represents the visual prior of input image, P denotes the textual prompt describing the image content, and I^{gen} is the output image. However, controllable models often overlook input image distributions due to gaps between training data and target domains, leading to distribution shifts in generated images. To address this, we integrate an Image-to-Image (Img2Img) mechanism into the Controllable Diffusion Model framework. This extends the function $\mathbf{G}_0$ to $\mathbf{G}_1 : I \times V \times P \to I^{gen}$, where the additional input I represents the reference image.

This approach ensures that the generated images not only maintain diversity at a moderate level but also exhibit improved similarity to the reference image, thereby achieving better alignment with the target data distribution. Furthermore, the Img2Img mechanism preserves the reference image's structural composition more effectively than traditional Text-to-Image (T2I) methods, as it utilizes the input image as a foundational guide during the diffusion process. As illustrated in Fig. 3a, our proposed pipeline incorporates two proposed methods to generate new image I^{gen}: *class-aware prompting* and *visual prior blending*, which generate P^* and V^*, respectively. These components are then combined with input image, enabling the Img2Img Controllable Diffusion Model to produce highly diverse images while preserving fine-grained details and maintaining the distributional characteristics of the original data.

$$I^{gen} = \mathbf{G}_1^*(I, V^*, P^*) \tag{2}$$

The overall process is depicted in Fig. 3b, highlighting the model's ability to produce highly diverse images while preserving fine-grained details and maintaining the distributional characteristics of the original data.

3.3   Controllable-Inpainting Diffusion Model

The issue of data generation out of the original domain when using controllable diffusion models has been highlighted in previous research [11]. This phenomenon can lead to a decline in model training performance as the dataset size increases. Although Sec. 3.2 introduces the Image-to-Image Controllable Dif-
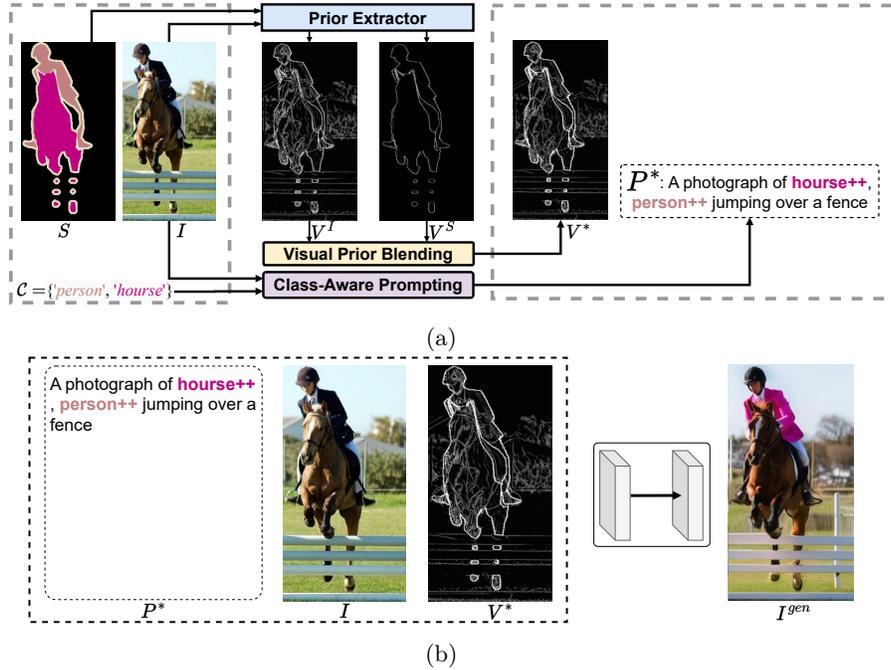
Fig. 3: Image generation using the Img2Img Controllable Diffusion Model.

fusion Model to mitigate this limitation, the transformation of the entire image makes it challenging to preserve the original data distribution. Therefore, we proposed to maintain the original image characteristics by employing an Inpainting Diffusion Model to modify specific regions of the image instead of transforming the entire image. To balance the advantages of both methods for the data augmentation task, we combine the Inpainting Diffusion Model with the proposed Img2Img Controllable Diffusion Model in Sec. 3.2, aiming to balance data diversity and the reliability of the generated images. The transformation function can represent the process of the Inpainting Diffusion model $\mathbf{G} : I \times M \times P \rightarrow I^{gen}$. Here, I is the input image, M is the mask specifying the regions to be modified, P represents the visual prior controlling the inpainting process, and I^{gen} is the generated image after the inpainting process.

[20] has been noted that relying on a pre-existing Inpainting Diffusion Model $\mathbf{G}$ does not ensure the newly generated objects conform to the mask M. Additionally, it does not guarantee that the generated objects retain the original objects. To address this limitation, we integrate the Controllable Model with the Inpainting Diffusion Model, resulting in a novel framework termed the Controllable Inpainting Diffusion Model. Precisely, controllable models such as T2I-Adapter [19] and ControlNet [18] extract visually structured information to inject into the Unet architecture of the Diffusion Model. In this approach, the generation process is conditioned not only on the mask M but also on the visual prior information V. We also utilize *class-aware prompting* and *visual prior*
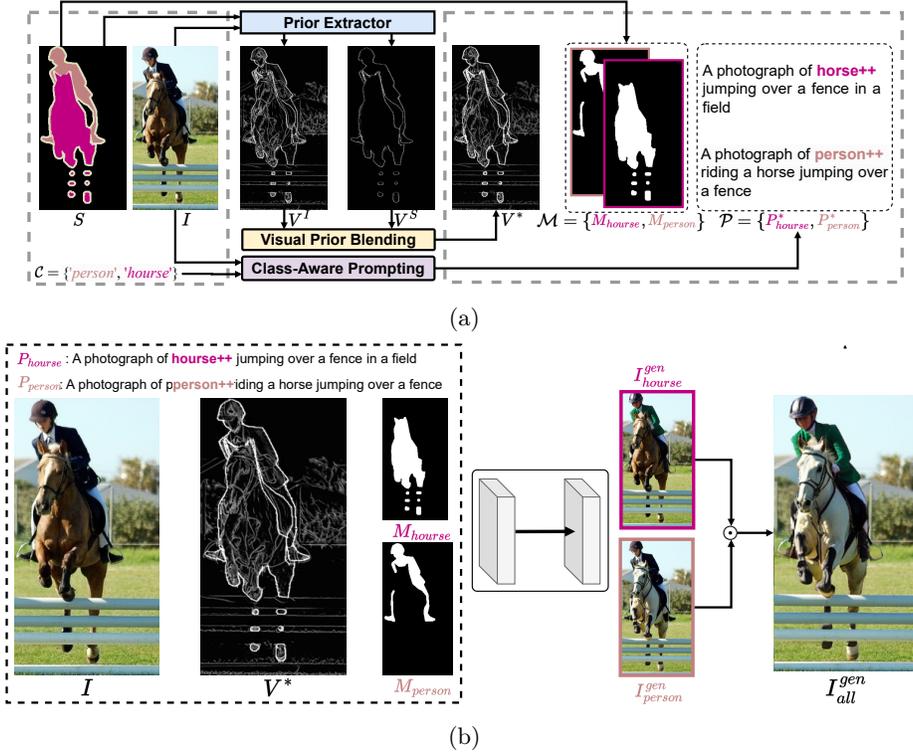
Fig. 4: Image generation using the Controllable Inpainting Diffusion Model.

blending techniques to reduce the risk of small objects being removed and replaced with background elements or inaccurately generated shapes; this process is illustrated in Fig. 4a. These improvements tackle the limitations associated to extends the function $\mathbf{G}_2^*\colon I \times V^* \times M \times P^* \to I^{gen}$. However, unlike the approach using the Img2Img Controllable Diffusion Model, the generation of objects in the Controllable Inpainting Diffusion Model is performed sequentially for each object type. This approach allows for a more accurate generation of objects, particularly when dealing with objects that have similar shapes. In this case, the images for each new object are generated as follows:

$$I_i^{gen} = \mathbf{G}_2^*(I, V^*, M_i, P_i^*) \tag{3}$$

where $M_i \in \mathcal{M}$ is the segmentation mask for the i-th class ($c_i \in \mathcal{C}$), determined by leveraging segmentation labels for each class. Meanwhile, $P_i^* \in \mathcal{P}^*$ is the prompt describing the data generation process for the class c_i. After obtaining the list of images $\{I_i\}$ generated based on the information for each class, we perform an image merging operation to produce the final composite image I_{all}^{gen}. This process relies on the list of masks $\{M_i\}$ to obtain an image with the labeled objects modified. This process is shown in Fig. 4b.

$$I_{all}^{gen} = I \odot \left(1 - \sum_{i=1}^{N} M_i\right) + \sum_{i=1}^{N} \left(I_i^{gen} \odot M_i\right) \tag{4}$$

4 Experiments

4.1 Datesets and implementation details

Datesets: We evaluate our synthetic data generation framework on two benchmark datasets: PASCAL VOC [3] (VOC07 and VOC12) and BDD100K [4]. To assess performance under data-limited scenarios, we conduct experiments on both the full VOC12 dataset (1,464 images), and its subsets [23]. Beyond standard object segmentation tasks, we further validate the model's capability to generate images under diverse environmental conditions (e.g., weather, scene) using BDD100K, a large-scale dataset capturing real-world driving scenarios for drivable area and lane segmentation tasks.

Implementation details: For object segmentation evaluation, we employ Deep -LabV3+ [24] (with ResNet50/101 backbones) and Mask2Former [25] (Swin-B backbone) implemented in the MMSegmentation framework, training for 30K iterations at 512×512 resolution with batch size 16 using AdamW optimization and default augmentations. For weather-conditioned segmentation on BDD100K [4], we adopt TwinLiteNet [26] for simultaneous lane and drivable area segmentation. Our image generation pipeline leverages SD-XL [17] controlled via T2I-Adapter [19] with Line Art as visual priors. The coefficient α in the *visual prior blending* method is set to 0.8.

Table 1: Comparison of mIoU (%) on the validation set when training models on the original dataset ($\mathcal{D}_0$) and when merging it with the synthetic dataset ($\mathcal{D}_0 \cup$ [11] vs. $\mathcal{D}_0 \cup$ ours), using DeepLabV3+ and Mask2Former model architectures.

Dataset			VOC7	VOC12				
Number images			209	92	183	366	732	1464
DeepLabV3+	Resnet50	$\mathcal{D}_0$	63.75	48.19	58.44	65.84	70.55	72.19
		$\mathcal{D}_0 \cup$ [11]	64.02	51.83	59.37	65.98	69.14	72.16
		$\mathcal{D}_0 \cup$ ours	**64.47**	**53.67**	**59.98**	**67.52**	**71.06**	**72.96**
	Resnet101	$\mathcal{D}_0$	67.61	54.06	62.88	67.85	73.06	76.19
		$\mathcal{D}_0 \cup$ [11]	68.79	**56.01**	63.09	68.89	73.05	75.68
		$\mathcal{D}_0 \cup$ ours	**68.81**	55.93	**64.08**	**69.54**	**73.83**	**76.91**
Mask2Former	Swin-B	$\mathcal{D}_0$	76.19	59.11	74.39	75.21	79.02	81.78
		$\mathcal{D}_0 \cup$ [11]	77.01	**65.01**	**76.67**	77.10	79.87	81.86
		$\mathcal{D}_0 \cup$ ours	**78.52**	64.51	76.45	**77.84**	**81.08**	**82.88**

Table 2: The table shows the performance of the Mask2Former (Swin-B) model when trained on (1) the real dataset, (2) the synthetic dataset, and (3) fine-tuned on the real dataset after pre-training on the synthetic dataset. The compared methods include generating synthetic data with pseudo-labels [9,10,27] and synthetic data based on the original dataset [11]

	Real images		Synthetic images					mIoU (%)
	VOC (5k)	VOC (1,5k)	DiffuMask [10] (60k)	DD [9] (40k)	Attn2Mask [27]	SG [11] (1,5k)	Ours (2,9k)	
(1)	✓							83.4
		✓						81.8
(2)			✓					70.6
				✓				67.6
					✓			71.0
						✓		73.0
							✓	76.3
(3)	✓		✓					84.9
		✓		✓				82.4
		✓				✓		82.8
		✓					✓	84.0

4.2 Semantic segmentation result on VOC

To evaluate our proposed data augmentation method, we compare models (Deep-LabV3+ and Mask2Former) trained on the original dataset ($\mathcal{D}_0$) and our augmented dataset ($\mathcal{D}_0 \cup \mathcal{D}_1^{gen} \cup \mathcal{D}_2^{gen}$). We also compare our method with Stronger Guidance [11], re-implemented using our training settings. Notably, we did not apply the object filter [14] or class balancing algorithm [11] to focus solely on synthetic image quality. As shown in Tab. 1, our method consistently improves semantic segmentation performance across datasets and architectures. While our method outperforms the baseline (trained on $\mathcal{D}_0$) in all configurations, it occasionally underperforms compared to [11] on smaller datasets but not significantly (VOC12 with 92 images when trained on DeepLabV3 Resnet101 and

Table 3: Evaluation of multi-task segmentation model performance across different environmental conditions when integrating our method with real data via merging/fine-tuning.

Condition	Ours	Number	Lane Line		Drivable Area
			Accuracy (%)	IoU(%)	mIoU(%)
Foggy		130	56.9	4.5	72.6
	✓	390	**67.8** / 67.5	8.3 / **8.7**	79.9 / **80.4**
Tunnel		129	80.5	9.3	73.4
	✓	387	84.4 / **85.2**	**15.2** / 14.9	86.0 / **87.3**
Gas Station		27	48.9	0.2	67.1
	✓	81	62.6 / **63.7**	0.4 / **0.5**	70.5 / **72.1**

D_0 [15] D_1^{gen} D_2^{gen}

Table 4: Quantitative comparison (FID /CLIP Score (ViT-B/32)).

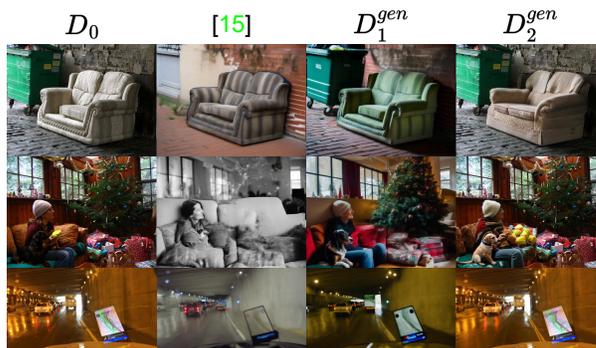|	CLIP ↑	FID ↓
[11]	0.81	114.49
$\mathcal{D}_1^{gen}$	0.84	101.92
$\mathcal{D}_2^{gen}$	0.92	72.22

Fig. 5: Some synthetic images generated using different methods.

VOC12 with 92 or 183 images when trained on Mask2Former). However, as the dataset size increases (e.g., VOC12 with 732 or 1464 images), our method consistently outperforms [11], which sometimes underperforms the baseline. This suggests that while highly diverse data improves accuracy on small datasets, it may cause distribution shifts that decrease performance as the number of samples in the dataset grows. Our method addresses this by balancing diversity and data consistency.

Additionally, we follow [10, 27], first training on synthetic data and then fine-tuning on real data (VOC12 with 1464 images), as synthetic data may not perfectly align with real data or domain shifts. Tab. 2 shows that our method achieves the highest mIoU (76.3%) when trained solely on synthetic data, outperforming other approaches. After fine-tuning, DiffuMask [10] achieves the best performance (84.9%), but requires 60k synthetic images for pre-training and 5k real images for fine-tuning. In contrast, our method achieves a competitive 84.0% mIoU with only 2.9k synthetic images and 1.5k real images, improving the baseline (81.8%) by 2.2%. This highlights the effectiveness of our approach.

4.3 Image generation based on environmental conditions

In addition to object segmentation, we evaluate the generation of images under different environmental conditions, such as fog, tunnel, and gas station scenarios. Using the TwinLiteNet model [26] for the drivable area and lane segmentation on the BDD100K dataset, we observe poor performance with fewer than 200 samples in these conditions. However, applying our method—through merging synthetic and real datasets or fine-tuning on real data—significantly improves the model's performance. This highlights our method's ability to generate synthetic data tailored to specific environmental conditions, enhancing model performance in real-world scenarios.

Table 5: Comparison of semantic segmentation performance (mIoU%) using different training strategies. Results are reported for Mask2Former models.

	$\mathcal{D}_0$	$\mathcal{D}_1^{gen}$	$\mathcal{D}_2^{gen}$	VOC7		VOC12	
				Number	mIoU (%)	Number	mIoU (%)
Train with Pure Real Data	✓			R: 209	76.19	R: 1464	81.8
Train with Pure Synthetic Data		✓		S: 209	74.32	R: 1464	74.22
			✓	S: 209	73.85	R: 1464	75.18
		✓	✓	S: 418	74.67	R: 2928	76.27
	✓	✓		R: 209	77.58	R: 1464	81.91
				S: 209	78.53	S: 1464	83.01
Merge with Real Data or Finetune on Real Data	✓		✓	R: 209	77.91	R: 1464	81.97
				S: 209	78.58	S: 1464	82.87
	✓	✓	✓	R: 209	78.52	R: 1464	82.88
				S: 418	80.21	S: 2928	84.02

4.4   Visualization and metrics for generated image quality

Qualitative results on the PASCAL VOC and BDD100K datasets, as illustrated in Figure 5, demonstrate that our method generates highly similar images to the original images (D_0). Specifically, images produced by the Img2Img Controllable Diffusion model (D_1^{gen}) and the Controllable Inpainting Diffusion Model (D_2^{gen}) not only exhibit diversity but also maintain structural similarity to the original images. In contrast, the method proposed in [11] yields inferior results, failing to preserve the structure and distribution of the generated images. In addition, we conducted quantitative evaluations using two metrics: FID and CLIP Score. The results in Tab. 4, evaluated on the VOC7 dataset, show that our method achieves higher scores, confirming its ability to generate high-quality images that closely align with the distribution of the original data.

4.5   Ablation Study

This section presents a comprehensive ablation study evaluating our method's components using the PASCAL VOC with Mask2Former.

Data Ablation Study: We evaluate the impact of synthetic data on semantic segmentation by comparing three strategies: (1) training on real data only, (2) training on synthetic data only, and (3) combining both. Results in Tab. 5 show that while training solely on synthetic data achieves notable accuracy, it is lower than training on real data. However, combining synthetic with real data, either by merging or fine-tuning real data after pre-training on synthetic data, significantly improves performance. The best results are achieved when fine-tuning on real data after synthetic pre-training and using multiple synthetic datasets ($\mathcal{D}_1^{gen}$ and $\mathcal{D}_2^{gen}$) further enhances performance, demonstrating that synthetic data is effective when combined with real data. These results confirm that while synthetic data cannot fully replace real data, it plays a key role in improving model performance, especially when merged or fine-tuned with real data.

Table 6: Performance of different text prompt selections when evaluating on VOC7 with Mask2Former (SwinB), The results are presented for model training using the merging mechanism.

Method	mIoU (%)
Simple text prompt	76.11
Generated caption	75.67
Class-prompt appending	77.81
Class-aware prompting	78.52

Table 7: Effect of increasing the number of synthetic images during training on the Mask2Former (SwinB) model. The values N_{real}/N_{syn} indicate the number of real and synthetic images, respectively.

VOC7		VOC12	
N_{real}/N_{syn}	mIoU(%)	N_{real}/N_{syn}	mIoU(%)
209/0	76.19	1464/0	81.8
209/418	78.52/80.21	1464/2928	82.88/84.02
209/836	79.21/81.53	1464/5856	82.01/85.33
209/1254	76.91/82.23	1464/8784	80.08/87.05

Text prompt selection: The performance of the model when selecting text prompts using different methods is detailed in Tab. 6. Our proposed *class-aware prompting* method demonstrates superior performance compared to previous methods. Specifically, our method achieves a performance of 78.52%. These results indicate that our text prompt generation method helps the model focus more effectively on the classes that need to be segmented, thereby significantly improving the performance of the semantic segmentation model.

Effect of different numbers of generated images in the synthetic data: In addition to generating two synthetic images per original image (via Controllable Inpainting Diffusion and Img2Img Controllable Diffusion), we conducted experiments by increasing the number of generated images to evaluate semantic segmentation performance. The results in Table 7 are presented in the format X/Y, where X denotes the performance from merging synthetic data with real data, and Y denotes the performance after fine-tuning on real data following pretraining. The results show that merging synthetic data with real data leads to degraded performance as the amount of synthetic data increases, whereas fine-tuning on real data after pretraining with synthetic data improves performance. These findings indicate that fine-tuning on real data after pretraining with a larger number of synthetic images can result in a more robust pre-trained model and improved overall performance.

5   Discussion and Conclusion

5.1   Limitations

Although our method shows promising results, there are several limitations. First, the quality of the synthesized images depends on the pre-trained generative model (SD). Second, generating high-quality synthetic images using diffusion models can be computationally expensive and time-consuming. Finally, while our method has potential for privacy-sensitive applications, this study only evaluates general datasets, so further validation is needed to ensure its effectiveness in specific scenarios.

5.2   Conclusion

In this work, we proposed a novel synthetic data augmentation pipeline that combines controllable diffusion models with advanced conditioning techniques to tackle the challenges of balancing diversity and reliability in semantic segmentation. Our method effectively generates high-quality synthetic data that preserves the structure of labeled objects and aligns well with real-world data distributions, demonstrating significant performance improvements on benchmark datasets like PASCAL VOC and BDD100K, particularly in data-scarce scenarios. Moreover, our approach effectively mitigates domain shift issues commonly associated with synthetic data generation, enabling more robust training.

Building on the success of image transformations guided by segmentation masks, we explore their potential for privacy protection applications. In privacy scenarios, sensitive regions identified by segmentation masks can be concealed using techniques like inpainting, ensuring privacy while preserving the quality of synthetic datasets for training segmentation models. Further exploration of these methods could enhance their applicability in privacy-sensitive domains.

6   Acknowledgement

References

1. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023.
2. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
3. M. Everingham, L. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *Int. J. Comput. Vision*, 2010.
4. F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *2020 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
5. M. Saran, F. Nar, and A. N. Saran, "Perlin random erasing for data augmentation," in *29th Signal Processing and Communications Applications Conference*, 2021.
6. Y. Chen, P. Zhang, Z. Li, Y. Li, X. Zhang, L. Qi, J. Sun, and J. Jia, "Dynamic scale training for object detection," 2021.
7. H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
8. W. Wu, Y. Zhao, H. Chen, Y. Gu, R. Zhao, Y. He, H. Zhou, M. Z. Shou, and C. Shen, "DatasetDM: Synthesizing data with perception annotations using diffusion models," in *Conference on Neural Information Processing Systems*, 2023.
9. Q. H. Nguyen, T. T. Vu, A. T. Tran, and K. Nguyen, "Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

10. W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models," *Proc. Int. Conf. Computer Vision (ICCV 2023)*, 2023.

11. Q. Che, D. Le, B. Pham, D. Lam, and V. Nguyen, "Enhanced generative data augmentation for semantic segmentation via stronger guidance," in *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, 2025.

12. C.-M. Feng, K. Yu, Y. Liu, S. A. Khan, and W. Zuo, "Diverse data augmentation with diffusions for effective test-time prompt tuning," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

13. B. Trabucco, K. Doherty, M. A. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," in *The Twelfth International Conference on Learning Representations*, 2024.

14. H. Fang, B. Han, S. Zhang, S. Zhou, C. Hu, and W.-M. Ye, "Data augmentation for object detection via controllable diffusion models," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

15. I. Goodfellow and et al., "Generative adversarial networks," *Commun. ACM*, 2020.

16. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

17. D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *International Conference on Learning Representations*, 2024.

18. L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *International Conference on Computer Vision (ICCV)*, 2023.

19. C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, "T2i-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Thirty-Eighth AAAI Conference on Artificial Intelligence*, 2024.

20. O. Kupyn and C. Rupprecht, "Dataset enhancement with instance-level augmentations," in *European Conference on Computer Vision*, 2024.

21. J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

22. Damian0815, "Compel: A library for conditioning and weighting in prompt-based models," 2023. [Online]. Available: https://github.com/damian0815/compel

23. Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao, and X. Le, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

24. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Computer Vision – ECCV 2018*, 2018.

25. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

26. Q.-H. Che, D.-P. Nguyen, M.-Q. Pham, and D.-K. Lam, "Twinlitenet: An efficient and lightweight model for driveable area and lane segmentation in self-driving cars," in *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2023.

27. R. Yoshihashi, Y. Otsuka, K. Doi, T. Tanaka, and H. Kataoka, "Exploring limits of diffusion-synthetic training with weakly supervised semantic segmentation," in *17th Asian Conference on Computer Vision*, 2024.