Keeping Segment Mask Quality with Self-generated Masks

Shinichi Mae[1], Ryosuke Yamada[2], and Hirokatsu Kataoka[2]

[1] TICO-AIST Cooperative Research Laboratory for Advanced Logistics(ALlab)
[2] National Institute of Advanced Industrial Science and Technology(AIST)

Abstract. Recent advances in generative models enable the creation of high-fidelity synthetic images from text prompts. Using these images as training data for visual models holds promise for efficient learning in recognition tasks, such as segmentation, which require detailed annotation and are difficult to gather data for. However, training on synthetic data might lead to model collapse in visual models. Effective methods are required to prevent this and achieve efficient learning with synthetic data. This study focuses on semantic segmentation and investigates efficient learning techniques using synthetic data generated by generative models. Specifically, we propose (i) a mask filtering method utilizing a segmentation model and (ii) a pre-training method named SemSegFDSL, which employs a dataset based on mathematical equations to construct visual models with a limited amount of synthetic data. Experimental results on the Pascal VOC dataset demonstrate that our method improves performance by 7.8% while using only half the synthetic data required by previous methods. These findings suggest that the quality of teacher labels affects model collapse and that filtering based on teacher labels and pre-training methods with synthetic data can effectively prevent this issue.

1 Introduction

In computer vision, large-scale image datasets, especially ImageNet [5], have revealed that the pre-trained models enable the assignment of multi-purpose recognition tasks. However, construction for a large-scale image dataset is highly expensive and labor-intensive, involving curation and manual assignment of teacher labels, which renders scaling the dataset challenging. In particular, segmentation datasets require rich annotation at the pixel level, which is reported to take approximately 90 minutes per image in Cityscapes [4]. Therefore, training visual models through efficient data generation is crucial for segmentation tasks that require rich annotation.

Conversely, large-scale generative models such as Stable Diffusion(SD) [12], which have recently trained on hundreds of millions of text–image pair datasets such as LAION [13], have achieved the generation of conditionally faithful and realistic synthetic images from arbitrary text prompts. These synthetic images can then be used as training data for visual models, considerably reducing the manual annotation effort and enabling efficient learning [16, 17]. For example,

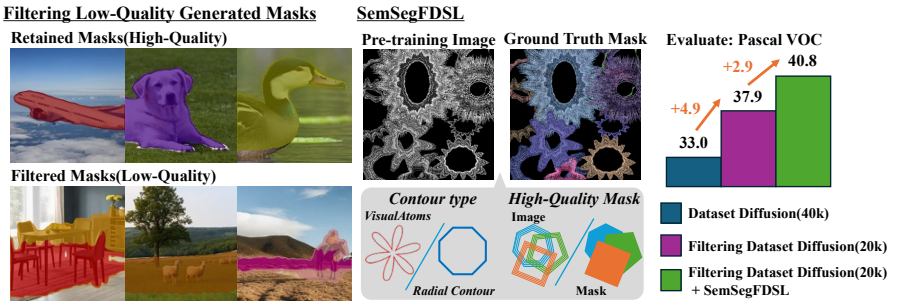


Fig. 1: Model performance is improved by filtering low-quality masks generated by the generative model and incorporating SemSegFDSL pre-training.

Dataset Diffusion [11] proposes a new framework for constructing high-quality segmentation datasets using SD. It introduces techniques such as adding class prompts, improving the accuracy of cross-attention, and indexing self-attention to generate realistic scenes with multiple objects and corresponding segmentation masks. Learning with synthetic data is thus very promising, particularly for recognition tasks where real data are difficult to collect and require rich supervised labels.

However, training on synthetic data generated by generative models may induce model collapse issues. Hataya et al. reported that when images generated by SD are mixed into real image datasets such as ImageNet and COCO, the generated images exhibit less variation compared to real images and include unnatural artifacts, which degrade the performance of the models [8]. Even in generative models for segmentation, the generated masks often contain errors, where regions that do not correspond to actual objects are mistakenly generated as masks.

Therefore, this study focuses on the quality of masks generated by generative models for segmentation and the performance of models trained on this data. We conduct exploratory experiments on learning with synthetic data generated from generative models to clarify the problems caused by synthetic data and avenues for their utilization. Specifically, as shown in Fig. 1, we propose (1) a mask filtering method using a segmentation model and (2) a pre-training method named SemSegFDSL, which uses a dataset based on mathematical equations to construct a visual model with limited synthetic data. The mask filtering method utilizes Segment Anything [10], which can predict highly accurate masks with zero-shot to filter out data with low-quality masks in synthetic datasets. SemSegFDSL constructs a highly accurate model with a limited amount of synthetic data by pre-training with completely accurate segmentation data, using figures drawn from mathematical formulas as input images and the contours of the figures as ground truth masks. Applying the proposed filtering method to a dataset of 40k samples generated by Dataset Diffusion, we observed a 4.9% improvement in model performance using only 20k filtered samples. Additionally, by using the proposed SemSegFDSL method for pre-training, we observed a further 2.9% improvement in performance.

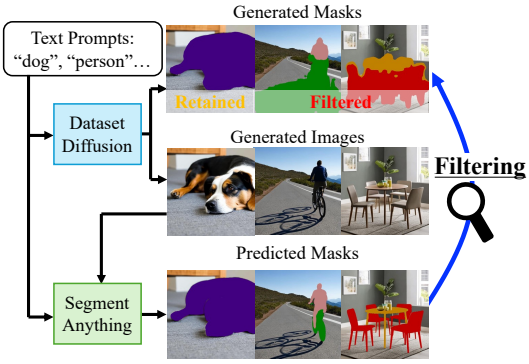


Fig. 2: Filtering method: We filter low-quality generated masks using SAM output masks.

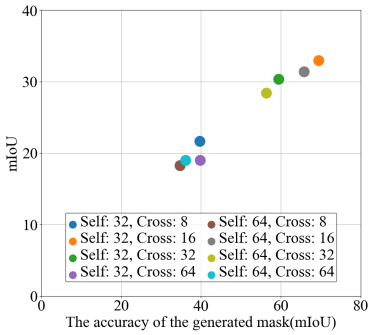


Fig. 3: Generated mask quality vs segmentation performance.

2 Method

This section presents efficient learning techniques using synthetic data generated by a generative model and a mathematical formula. First, Sec. 2.1 investigates the correlation between the quality of masks generated by the generative model and the performance of the segmentation model trained on this data as a pilot study. Next, Sec. 2.2 explains the filtering method to remove low-quality generated masks from synthetic training data. Finally, Sec. 2.3 describes the synthetic pre-training method to improve segmentation models under limited synthetic training data.

2.1 Pilot Study: Limitation of Generative Segmentation Masks

This experiment aims to reveal the dark side (*e.g.* model collapse) of training synthetic data by generative models. Specifically, we verify the correlation between the quality of the masks generated by the segmentation model and the recognition performance. We used Dataset Diffusion [11] to create synthetic datasets with various mask qualities, which generate synthetic images and segmentation masks from text prompts. Dataset Diffusion generates segmentation masks based on two parameters of the self-attention map and the cross-attention map, and the quality of the segmentation masks fluctuates with these two parameters. Thus, this experiment constructed a synthetic dataset of different mask qualities from 8 combinations of the self-attention map: {32, 64} and the cross-attention map: {8, 16, 32, 64}. All synthetic datasets consist of 40k images, 20 object classes, and one background class referencing Pascal VOC 2012(VOC) [6]. We refer to these synthetic datasets as synth-VOC.

We compare eight synthetic datasets in terms of mask quality and segmentation performance. For mask quality, we generate 1000 ground truth data from Segment Anything Models (SAM) [10] and human cross-checking, and we use mIoU calculated from the ground truth and generated data. For segmentation performance, we use Mask2Former [2] (Backbone Network: ResNet50 [9]) as a segmentation model. Mask2Former trains the generated synthetic dataset and calculates mIoU on 1449 VOC validation images. Fig. 3 shows a positive correlation between the quality of the generated mask and segmentation performance.

In other words, synthetic datasets with high-quality generated masks tend to achieve higher segmentation accuracy, while synthetic datasets with low-quality generated masks tend to exhibit lower segmentation accuracy. This result suggests that inaccurate supervised data from the generative model, as Dataset Diffusion, can lead to model collapse.

2.2 Filtering Low-Quality Generated Masks

We propose a filtering method for segmentation masks using SAM, a zero-shot prompt segmentation model, because recognition accuracy deteriorates when the quality of the generated mask is low based on the experimental results in Sec. 2.1. Our filtering method aims to remove noise data that negatively affects training and achieve data efficient learning. An overview illustration of the filtering method is shown in Fig. 2.

Specifically, we first input the synthetic image generated from Dataset Diffusion into the SAM to infer the segmentation masks. The synthetic image and its corresponding class name are input to SAM as a text prompt. Then, we compute the IoU for the masks generated by Dataset Diffusion and the output masks from SAM. Here, we use the synth-VOC (40k) dataset with the self-attention map: 32 and the cross-attention map: 16, as it achieved the highest model performance in Sec. 2.1. Lastly, the filtered synthetic dataset is constructed by selecting the Top-K (K = 10k, 20k, 30k) with the highest IoU. In other words, synthetic images with high IoU have many image regions that match the SAM output mask, so selecting them means using high-quality generated masks as training images.

2.3 Semantic Segmentation Formula-Driven Supervised Learning

This section introduces SemSegFDSL for the semantic segmentation task, where the pre-training dataset is automatically constructed based on a mathematical formula to achieve accurate recognition performance with limited synthetic data. FDSL is a synthetic pre-training method that automatically constructs a pre-training dataset by drawing a geometric shape in an image based on a mathematical formula to achieve supervised learning. For instance, Shinoda et al., [14] proposed Segmentation Radial Contour DataBase (SegRCDB) as the pre-training dataset for semantic segmentation. SegRCDB can automatically construct by drawing multiple objects with complex contours on the image and providing supervised labels at the pixel level from a mathematical formula. Despite synthetic images, the SegRCDB pre-trained model achieved comparable or better performance than the COCO-stuff [1] pre-trained model in Cityscapes [4]. In addition, Takashima et al. proposed Visual Atoms (VA) [15], a pre-training dataset consisting of synthetic images of objects with more complicated contours, and outperformed ImageNet in ViT pre-training.

Thus, we construct SemSegFDSL by improving SegRCDB. Specifically, we improve by adding complex geometric objects based on VA sinusoids as segmentation masks. SemSegFDSL consists of 20k pre-trained images, all of which are assigned a highly accurate segmentation mask based on a mathematical formula. A single image contains 32 objects, each corresponding to one of 255 categories defined by a mathematical formula. All other generation hyper-parameters are consistent with SegRCDB's baseline settings. Please refer to VA and SegRCDB for object category definition methods.

Table 1: Comparison of mIoU on Pascal VOC val between randomly selected masks generated by Dataset Diffusion and data filtered for low-quality masks using SAM.

Segmenter	Backbone	Dataset (data type)	#data	Filter	Iters	mIoU@VOC val
Mask2Former	ResNet50	SBD (real)	12k	-	133k	52.24
		Dataset Diffusion synth-VOC (synthetic)	10k	-	63k	26.35
				✓		**30.88**
			20k	-	125k	31.80
				✓		**37.93**
			30k	-	188k	33.79
				✓		**37.61**
			40k	-	250k	32.96

3 Experiments

We conduct experiments to validate the effectiveness of our filtering method for data-efficient fine-tuning and SemSegFDSL for pre-training Mask2Former. We first examine the segmentation performance in comparison with varios synthetic datasets based on filtering method in Sec 3.1. In addition, we explore the pre-training effect for the synthetic fine-tuning dataset by changing the parameters of SemSegFDSL, such as counter type and noise mask in Sec. 3.2. The training for each dataset using Mask2Former was conducted within the MMSegmentation framework [3], with the number of iterations adjusted according to the dataset size. All other hyper-parameters remained consistent with the official training settings and were not modified.

3.1 Effect of Filtering Low-Quality Masks (see Tab. 1)

In this experiment, we assess the effectiveness of our filtering method on a synthetic dataset. In the experimental setup, the synth-VOC and SBD datasets (consisting of 12,046 images as an extension of Pascal VOC) [7] are used for training, while the VOC validation set (VOC val) is used for evaluation. We apply our filtering method to synth-VOC (40k) and select the top 10k, 20k, and 30k data based on their IoU with the SAM mask. We also randomly selected an equivalent number of data for comparison. In Tab. 1, we compare the effectiveness of selecting synthetic images by filtering versus randomly sampling data. Firstly, Tab. 1 confirms that the filtered synth-VOC has better recognition accuracy than the randomly sampled synth-VOC in an equal number of data (10k, 20k, and 30k). Surprisingly, the synth-VOC (with filters: 20k and 30k) resulted in higher mIoU than the original synth-VOC (40k). This result indicates that using a smaller quantity of highly accurate segmentation labels for training yields better recognition performance than using a larger quantity of less accurate segmentation labels. However, since the filtered synthetic data still falls short compared to the real dataset, SBD, the next chapter will explore whether pre-training with synthetic data can improve the model performance.

3.2 Effect of SemSegFDSL Pre-training (see Tab. 2 and 3)

In this experiment, we examined the effectiveness of SemSegFDSL for pre-training when fine-tuning with synthetic data. Additionally, as an ablation study,

Table 2: Contour type

Contour	RC	VA	RC+VA
mIoU	39.79	**40.83**	40.51

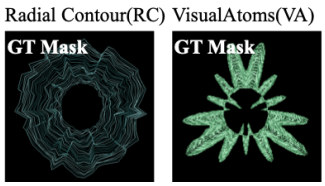
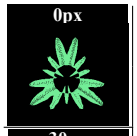
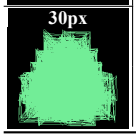


Table 3: Noise mask

Pre-training		Fine-tuning		mIoU VOC val
Noise	#data	Filter	#data	
	20k	-	40k	35.80
		✓	20k	**40.83**
		-	40k	35.40
		✓	20k	37.71

we investigated the impact of mask accuracy on model performance by introducing noise to the otherwise complete masks generated by SemSegFDSL, thereby degrading their quality in the pre-training data.

Effect of contour type. This experiment aims to investigate the pre-training effect of the newly added contour type. We explored the best effective contour type (RC, VA, RC+VA) in SemSegFDSL pre-training under the SemSegFDSL (20k) condition. We use the synth-VOC (with filter; 20k) as fine-tuning data and evaluate real images in VOC. Tab. 2 shows that the VA contour type provides the best results.

Effect of noise mask. In this experiment, we examined the impact of mask quality on model performance by introducing 30-pixel noise into the pre-training masks generated by SemSegFDSL. Additionally, for fine-tuning, we utilized data from synth-VOC (original, 40k) and synth-VOC (filtered, 20k) to investigate whether training on high-accuracy mask data can enhance model performance even with a reduced amount of synthetic data. As shown in the Tab 3, data with 30-pixel noise added to the ground truth masks in SemSegFDSL resulted in a performance degradation of up to 5.43% compared to noise-free data. These results demonstrate that higher mask accuracy in the training data, both during fine-tuning and pre-training, leads to improved model performance, even when using a limited amount of data.

4 Discussion and Conclusion

In this study, we quantitatively evaluated the degradation effect of low-quality masks generated by generative models on model performance and clarified the relationship between mask accuracy and model performance. We proposed a method for filtering low-quality masks using the Segment Anything Model (SAM) and demonstrated a 4.9% improvement in accuracy using only half of the conventional synthetic data. Additionally, by incorporating VisualAtoms (VA) as contour shapes for pre-training in SemSegFDSL, we achieved a 2.9% improvement in model performance when fine-tuning with a limited amount of synthetic data. This research highlights the potential degradation in model performance caused by AI-generated semantic segmentation masks. It underscores the need for improved segmentation data generation methods and the use of FDSL to enhance model performance, especially when training with AI-generated data.

References

1. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018) 4
2. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022) 3
3. Contributors, M.: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) 5
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 1, 4
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 1
6. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015) 3
7. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 international conference on computer vision. pp. 991–998. IEEE (2011) 5
8. Hataya, R., Bao, H., Arai, H.: Will large-scale generative models corrupt future datasets? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20555–20565 (2023) 2
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 3
10. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023) 2, 3
11. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 76872–76892. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/f2957e48240c1d90e62b303574871b47-Paper-Conference.pdf 2, 3
12. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1
13. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022) 1
14. Shinoda, R., Hayamizu, R., Nakashima, K., Inoue, N., Yokota, R., Kataoka, H.: Segrcdb: Semantic segmentation via formula-driven supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20054–20063 (October 2023) 4

15. Takashima, S., Hayamizu, R., Inoue, N., Kataoka, H., Yokota, R.: Visual atoms: Pre-training vision transformers with sinusoidal waves. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18579–18588 (June 2023) 4
16. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1206–1217 (October 2023) 1
17. Ye, H., Kuen, J., Liu, Q., Lin, Z., Price, B., Xu, D.: Seggen: Supercharging segmentation models with text2mask and mask2img synthesis. arXiv preprint arXiv:2311.03355 (2023) 1