A Generative Approach at the Instance-Level for Image Segmentation Under Limited Training Data Conditions (Student Abstract)

Thanh-Danh Nguyen[1, 2], **Vinh-Tiep Nguyen**[1, 2], **and Tam V. Nguyen**[3]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]University of Dayton, Dayton, OH 45469, United States
danhnt.ncs012024@grad.uit.edu.vn, tiepnv@uit.edu.vn, tamnguyen@udayton.edu

Abstract

High-accuracy image segmentation models require abundant training annotated data which is costly for pixel-level annotations. Our work addresses a high-cost manual annotating process or the lack of detailed annotations via a generative approach. In particular, our approach (1) proposes the conditional instance-level synthesis to enrich the limited data to enhance the segmentation performance, and (2) employs the generative architectures to complete the segmentation task under few-shot learning concepts. The initial results on the Cityscapes benchmark emphasize our potential generative solution on the instance segmentation task given limited data.

Introduction

Semantic scene understanding well supports downstream tasks and real-world applications including autonomous driving, medical imaging, and art. However, it is noted that the development of deep learning requires training such models on large abundant annotated data to achieve significant performance. This poses several problems related to the data, especially in cases when the data is scarce or the annotation process is high-cost. In this work, we focus on the intense limited training data concept of image segmentation by leveraging generation artificial intelligence (GenAI) to empower the task. Existing work (Nguyen et al. 2023) has successfully demonstrated the aforementioned idea in semantic segmentation, meanwhile, the instance segmentation domain remains challenging.

The contribution of this proposed work is three-fold. *First*, we propose a synthesis-based approach for training samples at the instance level given the pre-defined conditions to enrich the data for the instance segmentation task. *Second*, we propose utilizing the generative architecture to address the segmentation task under the few-shot learning concept. *Third*, considering the efficiency of the model, we employ an optimization technique given the synthesized data.

Proposed Approach

Strategy. To overcome the limited in both the number and diversity of the training data for the segmentation task, we leverage the power of GenAI. Recently, the image generation approach has achieved significant outcomes in synthesizing images given complicated conditions of text prompts or referenced images. Such methods based on the Diffusion model are well-trained with large data, i.e. Blended-Diff (Avrahami, Lischinski, and Fried 2022), DiffInpainting (Rombach et al. 2022), and GLIGEN (Li et al. 2023), are among the finest. The question is, how can we leverage such embedded huge knowledge in the GenAI models to support the downstream discriminative task of instance segmentation? We provide a detailed explanation of our approach to solving this problem as below.

Instance-level Synthesis Approach. Given image I and its corresponding instance annotation mask m, we first extract a list of captured instances to serve the instance-level augmentation. Accordingly, each input $inst$ consisting of an instance image X^i and its mask y^i, similarly formulated to n extracted instances. As illustrated in Figure 1-(a), an Instance-wise Image Generator Θ_G is responsible for synthesizing various versions $\hat{inst}$ of the input instance given the annotated information along with the class-guided prompt p formed on top of the instance category. The annotation mask of each instance is made used of for the training process of the Image Segmentation Model Θ_S to output the final predicted mask m. In this way, the Θ_G is under a plug-and-play manner to support different Diffusion-based architectures. This approach allows the generation model Θ_G to focus on the specific instance rather than the noisy background.

Embed Generative Architectures to Directly Segment Instances. Facing the drawback of the discriminative model addressing image segmentation including inductive bias when initiating anchors or bounding boxes, recent work (Gu, Chen, and Xu 2024) employs a generative approach to solve the discriminative task of image detection and segmentation. We inherit the denoising mechanism of the prior work (Gu, Chen, and Xu 2024) which introduces the noisy factor into the training samples, then propose a decoder to predict the outputs. Instead of using the synthesized results $\hat{inst}$, we directly embed an Extractor $e(\Theta_G)$ originated from the original generative model Θ_G to enrich the model Θ_S. Thus, Θ_S skips the intermediate stages to directly utilize the knowledge of the large model Θ_G (as illustrated in Figure 1-(b)).

Few-shot Learning and Model Optimization. When em-

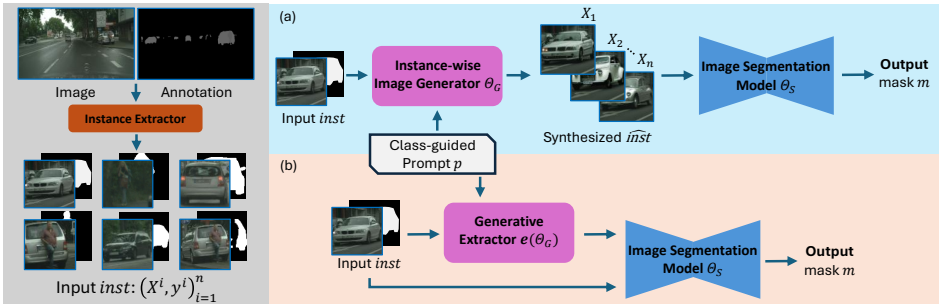

Figure 1: Our proposals on utilizing the GenAI to support the instance segmentation task via (a) instance synthesis approach and (b) generative architecture design.

Method	CLIPScore ↑	FID ↓
GLIGEN	0.79	40.65
DiffInpainting	0.81	31.03
BlendedDiff	**0.87**	**16.28**

*The best results are marked in **bold**.*

Table 1: Initial qualitative evaluation on Diffusion-based generation models

Method	Backbone	Synthesis	AP	AP50
FastInst	R50-FPN*		27.65	49.21
	R50-FPN	✓	**36.52**	**62.21**
OneFormer	ConvNext-L*		21.75	40.94
	ConvNext-L	✓	**38.93**	**64.91**
	Swin-L*		25.68	45.90
	Swin-L	✓	**35.75**	**61.01**

** denotes our reproduced experiments with hardware adaptation. Improved results are marked in **bold**.*

Table 2: Our initial instance segmentation results with the support of the proposed BlendedDiff-based approach.

ploying the generative models to segment images, we put the generative extractor $e(\Theta_G)$ into the segmentation track to explore the strength of the generative pre-trained network on large data. However, this poses an issue regarding the efficiency of the process. Thus, we first utilize few-shot learning to train the model with a few novel samples while inheriting base data knowledge. Then, we try optimizing the architecture of the generative model Θ_G in terms of sizes and parameters.

Preliminary Results

Initial Settings. Under the conditions of limited data, we focus on types of data which is abundant in the number of images but lack detailed-level annotations. To this end, we first initially examine our proposals on Cityscapes benchmark for urban scenes with approx. $2.5K$ annotated images to verify the mIoU or mAP. To report indirectly the image quality, we select the common CLIPScore and FID metrics. We plan to report the results on camouflaged object datasets and with more general image quality metrics in future experiments.

Baseline Initial Results. We evaluate several recent generation methods to guarantee the effectiveness of our proposed method on the Cityscapes dataset. In Table 1, we select BlendedDiff to report the initial results of utilizing the generative models to empower the recent image segmentation models. FastInst (He et al. 2023) and OneFormer (Jain et al. 2023) are the SoTA selected base models. As shown in Table 2, the support of the generation models improves average 12.04% AP evaluated on the validation set of Cityscapes under our customized reproduced results experimented on four GeForce RTX 2080Ti GPUs.

Conclusion and Future Work

Our initial work shows promising results in utilizing the generative approach to address the image segmentation task under limited training data conditions. In future experiments, we plan to explore and verify two major points: (1) how effective the embedded generative model contributes to segmenting instances, and (2) how to optimize the large model to strengthen the efficiency of the proposed methods under few-shot settings.

Acknowledgements

Thanh-Danh Nguyen was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2024.TS.068.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended diffusion for text-driven editing of natural images. In *CVPR*.

Gu, Z.; Chen, H.; and Xu, Z. 2024. Diffusioninst: Diffusion model for instance segmentation. In *ICASSP*. IEEE.

He, J.; Li, P.; Geng, Y.; and Xie, X. 2023. FastInst: A simple query-based model for real-time instance segmentation. In *CVPR*.

Jain, J.; Li, J.; Chiu, M. T.; Hassani, A.; Orlov, N.; and Shi, H. 2023. Oneformer: One transformer to rule universal image segmentation. In *CVPR*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *CVPR*.

Nguyen, Q.; Vu, T.; Tran, A.; and Nguyen, K. 2023. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *NeurIPS*, 36.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.