Inpainting-Based Labeled Dataset Generation and Pluggable Segmentation Method for Cellular Images

Jimao Jiang
School of Intelligence science and Technology
Peking University
ID 2301213295
jimaojiang@stu.pku.edu.cn

Rongyu Zhu
School of Mathematical Science
Peking University
ID 2301110075
2301110075@stu.pku.edu.cn

Abstract

Biomedical instance segmentation is a fundamental prerequisite in the diagnosis of histopathology. Considering the complexity of histopathological images and the scarcity and imbalance of annotated labels, costly and inefficient labeling is still inevitable. In this project, we propose an inpainting-based synthetic labeled dataset generation method for images with dense instances. By leveraging generative inpainting models, our approach not only ensures the quality of generated datasets but also provides a versatile solution for label generation, both for partial and full labels. Further, by integrating segmentation networks with generative models during training, we can improve performance as well as robustness for the plugged-in downstream task network. Experiments verify that our approach, plugged with a U-Net segmentation network, improves the nuclei segmentation results and noise-resistant ability on MoNuSeg dataset. Our code is available at https://github.com/JimaoJIANG/inpainting_based_cell_images_augmentation.

Group Member Contributions

Jimao Jiang (ID 2301213295, 60% Contribution): Literature research, Generative model training and testing, Part of the segmentation network testing, Slides and report modification.

Rongyu Zhu (ID 2301110075, 40% Contribution): Literature research, Part of the segmentation network testing, Slides and report writing.

1 Introduction

Biomedical image segmentation[1] is an active task in the field of digital pathology, which involves the process of partitioning histopathological images into multiple segments based on tissue types or anatomical structures. Cellular segmentation is an essential step in cancer analysis, diagnosis, classification, and grading, surgical planning, and treatment evaluation.

Different from natural image datasets, the variety of segmentation instances makes pixel-level biomedical image segmentation extremely challenging. And the efforts in collecting reliably annotated data can hardly be negligible, which requires high expertise due to the intrinsic difficulty of visually interpreting these histopathological images. This complexity arises both from their intricate structures and inconsistent acquisition conditions. Further, collecting real labeled data can often lead to imbalanced datasets that are unrepresentative of the overall data distribution. Therefore, the performance of most pixel-level segmentation tasks suffers due to the limited availability and imbalanced distribution of annotated training data. Reliable and efficient automatic segmentation methods are needed to provide accurate and consistent augmented data.

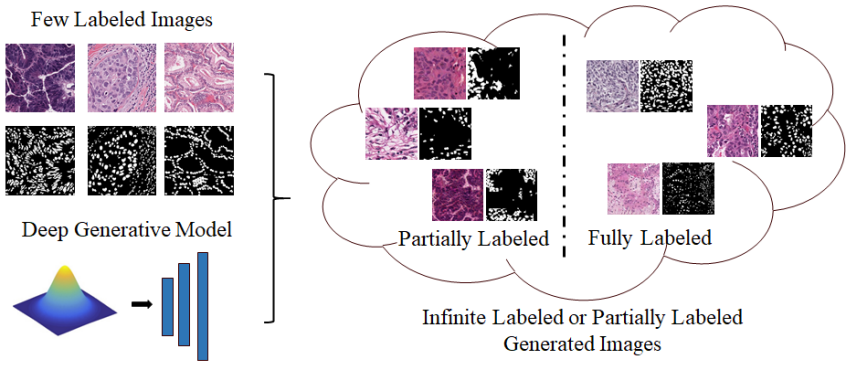

Figure 1: The project aims at using a small number of existing labeled images and a generative model to produce infinitely partially or fully labeled images.

In this project, we focus on generating labeled synthetic data for nuclei segmentation[3], an important branch of biomedical image segmentation. To address the challenges mentioned above, we exploit deep generative models to enrich the dataset to improve the performance on nuclei segmentation tasks, as shown in Fig1. Specifically, considering the dense instances nature of cellular images, we present an inpainting-based data augmentation technique to simultaneously generate cellular images along with their corresponding partial or full segmentation labels.

Our key works can be concluded as two parts, 1) A joint generative data augmentation method that generates images with dense cellular instances along with their corresponding partial labels by improving an existing inpainting method designed for natural images. And 2) A joint generative method for generating both cellular images and full labels with plugged-in segmentation network, meanwhile enhancing the robustness of the corresponding segmentation network. Our key contributions include:

- We propose a universal inpainting-based method for generating labeled datasets for images with dense instances, addressing the challenges of labeled data scarcity and imbalance.
- The corrupted and preserved areas in inpainting-based methods naturally provide partial segmentation labels for partially supervised training, and further allow for the integration of aribitary segmentation network to enhance robustness.
- By enriching the training dataset with synthetic labeled data, we improve the performance and robustness of nuclei segmentation in biomedical image analysis on MoNuSeg dataset.

2 Related Works

Traditional Biomedical Image Segmentation Methods. Research in biomedical image segmentation can be traced back to the 1970s and 1980s, when computer vision and image processing techniques began to develop. Before the rise of deep learning techniques, biomedical image segmentation relied heavily on traditional computer vision and image processing techniques. These methods include threshold-based segmentation[7, 2], edge detection[6, 15], region growing[13, 16], watershed algorithms[17, 11], and graph cut[10, 9]. However, most of these techniques require manual parameterization and remain highly sensitive to image quality, contrast, and noise. Thus, traditional methods are often limited to process simple biomedical images with regular shapes, but their performance suffers backgrounds get complex or with fuzzy boundaries.

Deep Biomedical Image Segmentation Methods. The emergence of deep learning techniques has significantly advanced the field of biomedical image segmentation, and deep convolutional neural networks have begun to demonstrate superior performance in image segmentation tasks. We will only briefly describe the articles that are closely related to our work. The U-Net architecture[12], a deep learning model designed specifically for medical image segmentation, was proposed in 2015 and has quickly become a standard tool in this field. AC-Former[4] combines the consistency constraints of affine transformations and the Transformer architecture to improve the accuracy and robustness of detecting and categorizing different types of cell nuclei in complex biomedical images. Kato et al.[5] present a novel approach to cell image segmentation that is capable of handling one-shot learning

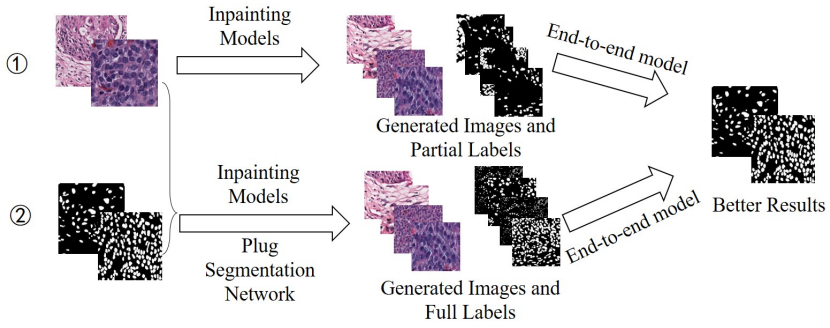

Figure 2: Main pipeline for two parts of our project.

scenarios and partially-supervised environments with a small set of visual prompts. However, the performance of these methods is limited by the scarcity of labeled data, due to the labor-intensive nature of biological data annotation.

Generative Methods in Data Augmentation. Generative methods are often used to do data augmentation to improve the generalization ability of machine learning models, especially when faced with data imbalance or data scarcity. cGANs[8] is a variant of Generative Adversarial Networks (GANs), it guides the generations process by using additional information such as class labels. ControlNet[20] can effectively control diffusion models such as Stable Diffusion to generate images that better match the user's mental expectations. CellGAN[14] introduces a generative adversarial network (GAN) framework that synthesizes realistic cervical cell images to augment the limited datasets used in cytopathological image classification, thereby improving the accuracy and robustness of diagnostic models. HandsOff[19] uses GANs to create synthetic datasets with labels, while DatasetDM[18] achieves this through diffusion models. However, these methods either cannot generate labels while producing images [8][20], or use a pretrained model to generate downstream labels [19][18]. In contrast, our method can obtain partial labels while generating images, or train a segmentation network simultaneously with the generation.

3 Methods

3.1 Motivation

Our goal is to generate labeled datasets for images with dense instances. In this project, we focus on nuclei segmentation for cellular images. The most notable characteristics that distinguish biomedical image datasets from natural images include: 1) small amounts of data; 2) dense instances; 3) diverse morphologies of the data (e.g., cells vary in size and length); 4) class imbalance.

Considering the characteristics of the dataset, we intend to exploit the data augmentation using deep generative models, 1) to generate new images with partial labels, or furthermore, 2) to generate new images with full labels, and eventually to be able to generate infinite labeled or partially labeled images. Our main idea can be concluded in Fig.2. We choose image inpainting methods for three reasons:

1. Firstly, cellular images contain dense instances and irregular data distributions. Unlike image inpainting methods for natural images, inpainting-based methods for cellular image generation can ensure continuous context along the edges of corrupted areas while filling the interior with variety due to the lack of clear inner context. This approach results in diverse outputs and can generate specified categories based on the preserved areas.
2. Secondly, although generative models are highly capable of synthesizing images within known data distributions, most of these methods focus solely on generating images, ignoring the joint generation of labels. Some existing methods therefore generate labeled synthetic datasets by using a separate pretrained model to label the generated images [19][18]. In contrast, our method can definitely obtain a partial label based on the corrupted area, as long as we generate an image, thus enabling partially supervised training.

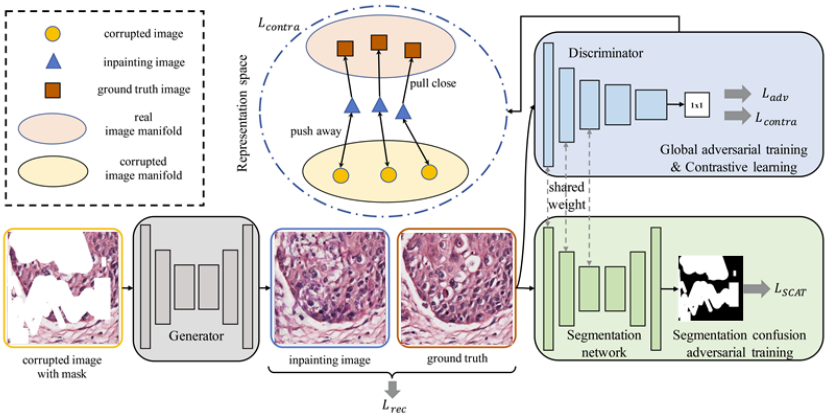

Figure 3: The overall framework of the proposed method for partially labeled image generation.

3. Thirdly, integrating the image segmentation network with generative learning facilitates both partially and fully supervised training. This approach exposes the network to a broader spectrum of images, including noisy and low-quality samples when the network has not yet been well optimized, thus enhancing the robustness of the segmentation network.

3.2 Project Part I-Partially labeled Image Generation with Inpainting Models

In this section, we present our first work, i.e., generating images with dense cellular instances along with their corresponding partial labels by improving an existing inpainting method [21] designed for natural images. We transfer the inpainting network designed for natural images and train the network with cellular images from scratch. During testing, we can control the specific category of images we need thanks to the preserved areas of the input image instead of random noise. Thus, we can generate diverse outputs with a specified category.

In addition, the proposed inpainting-based method can be utilized even in the absence of labels, as training under the current setting does not require input of labels. The generation results exhibit similarities to the existing controllable generation method, such as [8] and [20].

We use an adversarial training framework for image inpainting with segmentation confusion adversarial training (SCAT) and contrastive learning[21]. The framework is shown in Fig.3.

The input accommodates the original image along with any mask, enabling us to select any portion of the cellular image and prompt the network to generate new content. Operating within a GAN-like framework, the training occurs in a non-stationary environment. Following [21], we leverage the force of contrastive learning to prevent degraded or cyclic inpainting results. By ensuring that the repaired images progressively diverge from the corrupted ones in the representation space of the discriminator, the adversarial training can be stabilized and enhanced. Additionally, inspired by how humans recognize a low-quality repaired image, the segmentation network acts as another discriminator, distinguishing artificial segments within the image, thereby enhancing the overall performance.

The total loss of the network can be written as follow:

$$\mathcal{L}_{partial} = \lambda_{adv}(\mathcal{L}_{adv} + \mathcal{L}_{SCAT}) + \mathcal{L}_{contra} + \lambda_{rec}\mathcal{L}_{rec}, \tag{1}$$

and $\mathcal{L}_{adv}$ is the standard global adversarial training loss,

$$\mathcal{L}_{adv} = \min_{G}\max_{D} \mathbb{E}_{x}[\log D(x)] + \mathbb{E}_{\bar{x}}[\log(1 - D(\bar{x}))]. \tag{2}$$

The segmentation confusion adversarial training loss is defined as,

$$\mathcal{L}_{SCAT}(G) = -\mathbb{E}[\frac{1}{HW}\sum_{i=1}^{HW}[\bar{m}_i \log S(\bar{x})_i + (1-\bar{m}_i)\log(1-S(\bar{x}_i))]], \tag{3}$$

$$\begin{aligned}\mathcal{L}_{SCAT}(S) = -\mathbb{E}[&\frac{1}{HW}\sum_{i=1}^{HW}[m_i \log S(\bar{x})_i + (1-m_i)\log(1-S(\bar{x}_i))] \\ &+\frac{1}{HW}\sum_{i=1}^{HW}[\bar{m}_i \log S(\bar{x})_i + (1-\bar{m}_i)\log(1-S(\bar{x}_i))]],.\end{aligned} \tag{4}$$

$\mathcal{L}_{SCAT}(G)$ and $\mathcal{L}_{SCAT}(S)$ encourages G to produce better inpainting images such that S mistakes the generated regions in the inpainting image as valid regions ($\bar{m}$ refers to the mask filled with all 1s). The contrastive loss which stabilizes and improves the adversarial training is written as,

$$\mathcal{L}_{contra} = \lambda_{text}\mathcal{L}_{text} + \lambda_{sem}\mathcal{L}_{sem}. \tag{5}$$

We use the same definition of $\mathcal{L}_{text}$ and $\mathcal{L}_{sem}$ to control the push and pull forces as in [21].

Finally, the reconstruction loss is formulated as,

$$\mathcal{L}_{rec} = \mathbb{E}\|\hat{x} - x\|_1, \tag{6}$$

which directly describes the difference between the inpainting image and the ground truth.

Since we need variety when generating diverse cellular images with image inpainting, we reduce the weight of $\mathcal{L}_{rec}$ after training a certain number of steps (in experiments we replace λ_{rec} with $\lambda_{rec}/10$ after training $1/3$ of the total number of steps) to prevent the network from generating nearly identical images. In addition, we increase the mask ratio in the training phase to generate more samples, and decrease the mask ratio in the testing phase to retain enough label information for downstream task training.

After the generator is well trained, we can obtain a partial label based on the preserved area, as long as we generate an image, thus enabling partially supervised training.

As shown in Fig.4, we feed both the real images and generated images to the U-Net to do the partially supervised segmentation training. And the total loss is:

$$\mathcal{L}_{seg} = \mathcal{L}_{full} + \mathcal{L}_{part}, \tag{7}$$

where

$$\mathcal{L}_{full} = \text{BCELoss}(s, s_{pred}) \tag{8}$$

$$\mathcal{L}_{part} = \text{BCELoss}(s \odot (1-m), s_{pred} \odot (1-m)) + \lambda_{reg}\mathcal{L}_{reg} \tag{9}$$

correspond to the loss of original images with full labels and generated images with partial labels, respectively. And

$$\mathcal{L}_{reg} = \text{BCELoss}(s \odot m, \bar{m}) + \text{BCELoss}(s \odot m, 1-\bar{m}) \tag{10}$$

is the regularization term that prevents the segmentation within masked area to be all black or all white.

3.3 Project Part II-Fully labeled Image Generation with Plugged-in Segmentation Network

Instead of training the generator and segmentation network separately, we can embed the semi-supervised segmentation network into the generative network, since there is a constant flow of partially labeled generative images with fully labeled original images to the segmentation network. The network can expose to a larger range of images, including noise and low-quality samples in the early stages of network optimization. The segmentation network could become more robust and more noise-resistant compared to the independent training.

It is worth noting that, different from the approach in Section 3.2 that only uses the original image itself as input, the labels for training datasets are indispensable in this setting.

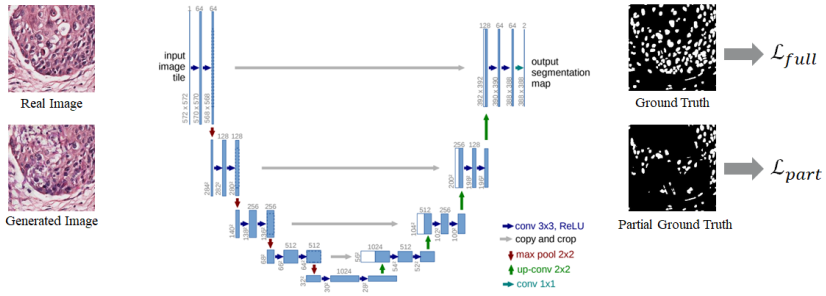

Figure 4: Downstream task. We use a simple U-Net for nuclei instance segmentation.

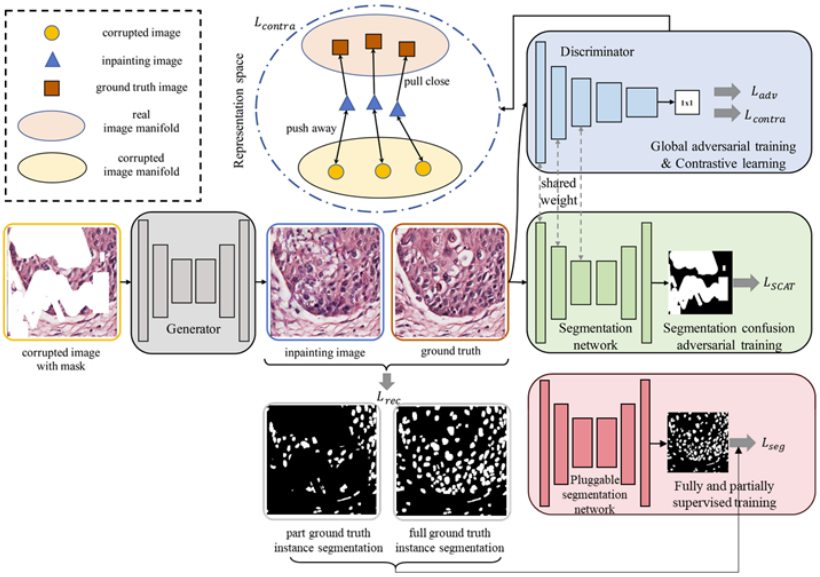

Figure 5: Fully labeled image generation with plugged-in segmentation network.

The plugged-in segmentation network is shown in Fig.5. The total loss of the network can be written as the sum of 1 and 7:

$$\begin{aligned}\mathcal{L}_{full} &= \alpha_{partial}\mathcal{L}_{partial} + \alpha_{seg}\mathcal{L}_{seg} \\ &= \alpha_{partial}\left(\lambda_{adv}(\mathcal{L}_{adv} + \mathcal{L}_{SCAT}) + \mathcal{L}_{contra} + \lambda_{rec}\mathcal{L}_{rec}\right) + \alpha_{seg}\left(\mathcal{L}_{full} + \mathcal{L}_{part}\right), \end{aligned} \quad (11)$$

where $\alpha_{partial}$ and α_{seg} are used to balance the generative and discriminative subtasks.

After training, we can either use the final segmentation network directly for downstream segmentation tasks or employ it to generate complete labels for all the generated images. This allows us to create a new fully labeled dataset, which can then be used to retrain a new U-Net independently. Both approaches will be evaluated in our experiments.

4 Experiments

4.1 Dataset and Metrics

Dataset. We use the MoNuSeg dataset to evaluate our method. MoNuSeg is a hypothetical dataset designed for monocular cell nucleus segmentation tasks, a critical component in computational pathology and biomedical image analysis. It provides a rich collection of annotated images, capturing various cell types and their nuclei with diverse morphologies and distributions. The dataset is curated to support the development and evaluation of advanced image segmentation algorithms, particularly those leveraging deep learning techniques for accurate and efficient cell nucleus detection and

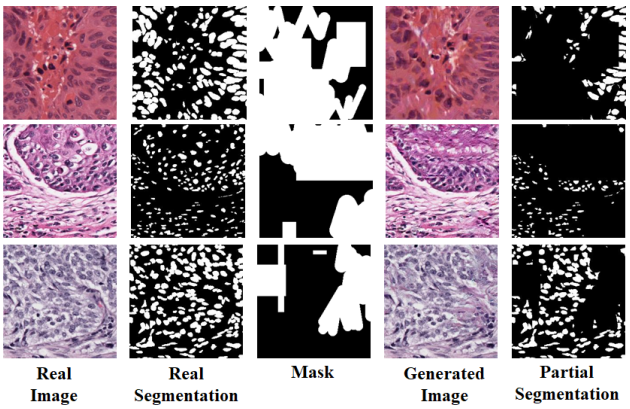

Figure 6: Partially labeled image generation with random masks.

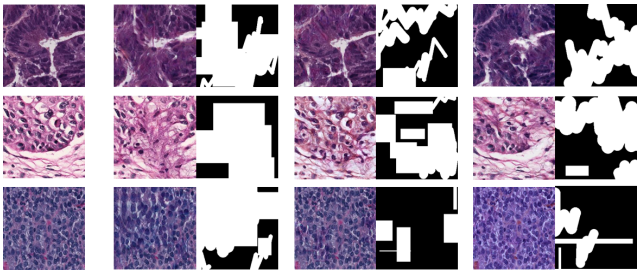

Figure 7: Controllable cellular image genration based on one single image.

classification. Training data of MoNuSeg containing 37 images and around 22,000 nuclear boundary annotations, testing data containing 14 images with additional 7000 nuclear boundary annotations.

Metric. We use the Dice Similarity Coefficient (DSC) to evaluate the segmentation performance of biomedical images.

$$\text{DSC}(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{12}$$

where $|A|$ and $|B|$ refer to the cardinality of sets A and B.

4.2 Implementation Details.

We implement the proposed method using the PyTorch toolkit on a PC with an NVIDIA GTX 3090 GPU. We use the Adam optimizer with a learning rate of 1e-4 at the beginning. The training takes 30,000 steps. Following [21], we set $\lambda_{adv} = \lambda_{text} = 1$, $\lambda_{sem} = \lambda_{rec} = 10$. In addition, we set $\alpha_{partial} = \alpha_{seg} = 1$. All the masks and images for training are of size 256×256.

4.3 Partially labeled Image Generation

Fig.6 shows that the inpainting network in Section 3.2 can generate different and also meaningful cellular images from the original with arbitrary mask, along with their partial labels. The controllability of generated images is shown in Fig.7, where we mask different parts of the same image to vary the generated results but constrain the results to be consistent with the preserved area.

4.4 Fully labeled Image Generation

With plugged-in segmentation network as depicted in Section 3.3, we could generate synthetic cellular images with full labels, as shown in Fig.8. The quality of the generated labels is natural enough to improve the performance of downstream tasks, as verified in Section 4.5.

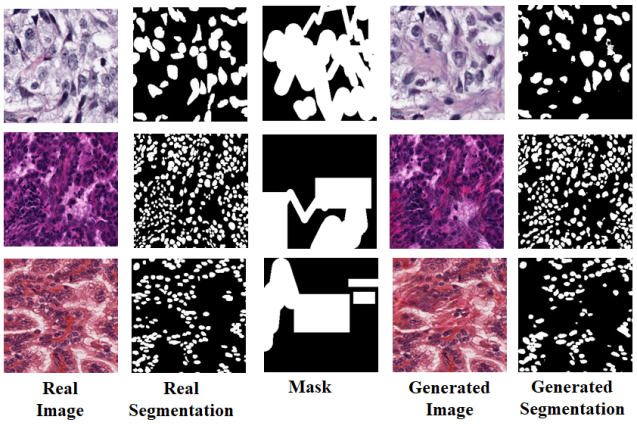

Figure 8: Fully labeled image generation using plugged-in segmentation network.

Baseline	Method	Data Augmentation	Supervision	Training Scheme	Random Noise Ratio				w/o Noise
					5%	10%	15%	20%	
U-Net	Vanilla	None	Full	Independent	0.745	0.744	0.709	0.664	0.746
	Vanilla+Aug	Basic	Full	Independent	0.781	0.767	0.753	0.741	0.799
	Generative-I	Basic+Generative	Full+Part	Independent	0.806	0.788	0.766	0.752	0.808
	Generative-II	Basic+Generative	Full+Part	Plugged-in	0.809	0.806	0.797	0.780	0.809
	Generative-III	Basic+Generative	Full+Full	Independent	0.825	0.815	0.791	0.759	0.826

Table 1: Quantitative evaluation of our generated images and corresponding labels regarding the DSCs on different training and noise settings. "Basic" refers to basic data augmentation strategy including random crop, random resize, random flip and color jitter.

4.5 Quantitative Analysis

The quantitative analysis is shown in Table.1. Generative-I refers to the method in Section 3.2. Generative II corresponds to evaluation by the plugged-in UNet as detailed in Section 3.3. Generative-III involves training a separate U-Net with the generated images and corresponding full labels, using the well-trained inpainting network and plugged-in U-Net from Generative-II.

The performance of the segmentation networks improve with generated labeled images, both for partial and fully labeled images, demonstrating the efficacy of generated labeled images in enhancing segmentation quality. The best result comes from Generative-III at 0.826, the method using fully labeled generated images to independently train a new segmentation network, making it the best overall method in terms of raw performance. On the other hand, in terms of robustness, Generative-II is the most noise-resistant, which confirms our previous statement. However, the performance of vanilla U-Net decreases significantly with increasing noise. Fig.9 shows the segmentation results of images with and without random noise. And Generative-III also maintains competitive performance with noise. Visually, Generative-II leads to the least amount of noise in segmentation images as the noise ratio increases.

4.6 Ablation study

To validate the role of generative model in our experiments, we conduct an ablation study as illustrated in Fig.10. We replace the corrupted area in cellular images with random noise, white and black, and compare the performance of segmentation network trained on these images with the one trained on inpainting images. The DSCs of generative is slightly higher than others. But the advantage becomes more obvious in qualitative results, as shown in Fig.10-Right, the generative segmentation has less tiny noisy dots and is more consistent with the ground-truth.

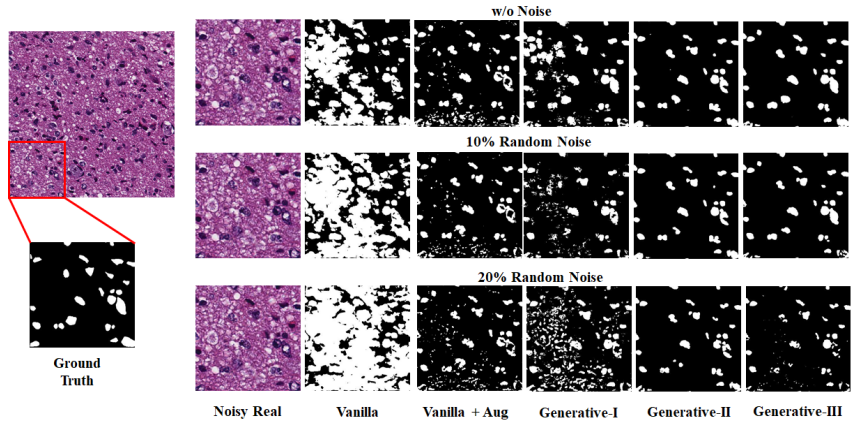

Figure 9: Visualization for image segmentation under different training and noise settings for both models with and without generative data.

Masked Area	DSCs
Random	0.785
White	0.786
Black	0.791
Generative	0.795

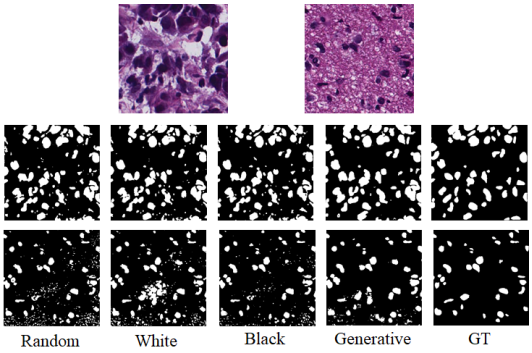

Figure 10: Comparison of segmentation results with different corrupted area.

5 Conclusion

In this work, we propose a novel inpainting-based labeled dataset generation method for biomedical images with dense instance. Our project consists of two main components. First, we adapt an inpainting-based network to generate new images with corresponding partial labels, which has addressed the problem of data scarcity and data imbalance, and demonstrating advantages in partially supervised learning. Second, we generate both histopathological images and full labels with a plugged-in segmentation network in a semi-supervision manner, meanwhile improving its robustness. The full-label output of plugged-in segmentation network can be used to independently retrain a new segmentation network with better results. Experiments show that the final segmentation network shows improved performance and noise resistance ability when trained with synthesized images and labels by our generative model.

Potential future work. Our method is not limited to the specific inpainting and segmentation method used in our report. In fact, our work supports any image inpainting network and any segmentation network. It would be interesting to investigate the performance with other inpainting and segmentation methods.

References

[1] Yahya Alzahrani and Boubakeur Boufama. Biomedical image segmentation: a survey. *SN Computer Science*, 2(4):310, 2021.

[2] Neeraj J Gadgil, Paul Salama, Kenneth W Dunn, and Edward J Delp. Nuclei segmentation of fluorescence microscopy images based on midpoint analysis and marked point process. In *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)*, pages 37–40. IEEE, 2016.

[3] Tomohiro Hayakawa, VB Surya Prasath, Hiroharu Kawanaka, Bruce J Aronow, and Shinji Tsuruoka. Computational nuclei segmentation methods in digital pathology: a survey. *Archives of Computational Methods in Engineering*, 28:1–13, 2021.

[4] Junjia Huang, Haofeng Li, Xiang Wan, and Guanbin Li. Affine-consistent transformer for multi-class cell nuclei detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21384–21393, 2023.

[5] Sota Kato and Kazuhiro Hotta. One-shot and partially-supervised cell image segmentation using small visual prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4294–4303, 2023.

[6] SS Kattire and AV Shah. Boundary detection algorithm implementation for medical images. *International Journal of Engineering Research & Technology (IJERT)*, 3(12):2278–0181, 2014.

[7] Cheng Lu, Muhammad Mahmood, Naresh Jha, and Mrinal Mandal. A robust automatic nuclei segmentation technique for quantitative histopathological image analysis. *Analytical and Quantitative Cytology and Histology*, 34:296–308, 2012.

[8] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[9] Tomoyuki Nagahashi, Hironobu Fujiyoshi, and Takeo Kanade. Image segmentation using iterated graph cuts based on multi-scale smoothing. In *Computer Vision–ACCV 2007: 8th Asian Conference on Computer Vision, Tokyo, Japan, November 18-22, 2007, Proceedings, Part II 8*, pages 806–816. Springer, 2007.

[10] Kaustav Nandy, Rama Chellappa, Amit Kumar, and Stephen J Lockett. Segmentation of nuclei from 3d microscopy images of tissue via graphcut optimization. *IEEE Journal of Selected Topics in Signal Processing*, 10(1):140–150, 2015.

[11] Uppada Rajyalakshmi, S Koteswara Rao, and K Satya Prasad. Supervised classification of breast cancer malignancy using integrated modified marker controlled watershed approach. In *2017 IEEE 7th International Advance Computing Conference (IACC)*, pages 584–589. IEEE, 2017.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[13] Brijesh N Shah, Satish K Shah, and YP Kosta. A seeded region growing algorithm for spot detection in medical image segmentation. In *2011 International Conference on image information processing*, pages 1–4. IEEE, 2011.

[14] Zhenrong Shen, Maosong Cao, Sheng Wang, Lichi Zhang, and Qian Wang. Cellgan: Conditional cervical cell synthesis for augmenting cytopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 487–496. Springer, 2023.

[15] Liyilei Su, Xianjun Fu, Xiaodong Zhang, Xiaoguang Cheng, Yimin Ma, Yungen Gan, and Qingmao Hu. Delineation of carpal bones from hand x-ray images through prior model, and integration of region-based and boundary-based segmentations. *IEEE Access*, 6:19993–20008, 2018.

[16] Shan Tan, Laquan Li, Wookjin Choi, Min Kyu Kang, Warren D D'Souza, and Wei Lu. Adaptive region-growing with maximum curvature strategy for tumor segmentation in 18f-fdg pet. *Physics in Medicine & Biology*, 62(13):5383, 2017.

[17] Mitko Veta, A Huisman, Max A Viergever, Paul J van Diest, and Josien PW Pluim. Marker-controlled watershed segmentation of nuclei in h&e stained breast cancer biopsy images. In *2011 IEEE international symposium on biomedical imaging: from nano to macro*, pages 618–621. IEEE, 2011.

[18] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 54683–54695. Curran Associates, Inc., 2023.

[19] Austin Xu, Mariya I. Vasileva, Achal Dave, and Arjun Seshadri. Handsoff: Labeled dataset generation with no additional human annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7991–8000, June 2023.

[20] 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 (ICCV)*, pages 3836–3847, October 2023.

[21] Zhiwen Zuo, Lei Zhao, Ailin Li, Zhizhong Wang, Zhanjie Zhang, Jiafu Chen, Wei Xing, and Dongming Lu. Generative image inpainting with segmentation confusion adversarial training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3888–3896, 2023.