Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France – August, 2025
Paper No. ICBES 139
DOI: 10.11159/icbes25.139

Supercharging Generalized Cell Segmentation with Diffusion Synthesized Data

Henry Hou
The Bishop's School
7607 La Jolla Blvd, La Jolla, CA 92037, USA
henry.hou.27@bishops.com

Abstract **-** Medical doctors and biological researchers often rely on imaging techniques to analyze tissue structures and cellular morphology. For example, pathologists use H&E-stained images for disease diagnosis and grading. Since cells are fundamental components in many imaging modalities, an automatic and accurate cell segmentation pipeline can greatly benefit biological research and medical diagnostics. In recent years, deep learning has significantly advanced cell segmentation, with foundation models (FMs) emerging as a promising approach. One such model, CellViT, based on the Vision Transformer (ViT) architecture, has gained attention for its strong generalization capabilities across different cell imaging modalities and datasets. In this study, we evaluate the accuracy and robustness of CellViT on the publicly available MoNuSeg dataset for cell segmentation. While the model produces reasonable results, its performance can be further improved. To enhance CellViT, we develop a fine-tuning pipeline that integrates both manually labeled cell images (with precise cell annotations) and synthetic images generated by a diffusion model, DiffInfinite (with pseudo annotations). Our pipeline incorporates key components, including FM fine-tuning, human-in-the-loop feedback for using synthetic images, and FM training on a combination of real labeled and pseudo-labeled generated images. Experimental results demonstrate significant improvements in segmentation accuracy and generalization when applying our method to CellViT. This study highlights the potential of enhancing foundation models through fine-tuning and synthetic data augmentation, paving the way for more robust and accurate biomedical image analysis.

Keywords**:** Cell Segmentation, Foundation Models, Vision Transformer, CellViT, Synthetic Data, Diffusion Models, Fine-Tuning, Medical Image Analysis

1. Introduction

Machine learning, particularly deep learning, has rapidly expanded into the medical field, significantly impacting medical image analysis [9]. The accurate segmentation of medical images is crucial for various clinical applications, including disease diagnosis and treatment monitoring [5]. However, developing robust segmentation models faces several challenges [9]. Manual segmentation, the gold standard for creating training data, is highly labor-intensive and requires domain expertise, making it expensive. Furthermore, the availability of high-quality annotated medical images is often limited due to patient privacy regulations and the inherent complexity of medical data. Consequently, existing segmentation models frequently struggle with accuracy and generalization due to data scarcity. Foundation models (FMs) represent a promising direction in addressing these challenges [11]. Recent advancements such as Segment Anything [12] and its medical adaptation, SAM-Med [14], demonstrate the growing potential of universal segmentation models across domains, highlighting the importance of developing robust and generalizable solutions for clinical applications. CellViT, a Vision Transformer (ViT)-based FM trained on manually segmented images from The Cancer Genome Atlas (TCGA) [2], has shown capabilities in segmenting cell images, identifying cell states, and recognizing tissue types [3]. Despite its strengths, the performance of CellViT and similar models still requires improvement to meet the demands of clinical adoption. This study aims to enhance the performance of CellViT through a novel fine-tuning strategy. Specifically, we investigate how integrating synthetic data generated by a diffusion model (DiffInfinite) [7] can improve CellViT's segmentation accuracy on the MoNuSeg dataset [5]. We explore a human-in-the-loop approach for selecting high-quality synthetic images and assess the impact of augmenting the training data with these pseudo-labeled synthetic images on model accuracy and potential clinical applicability. Our central research question is how to effectively leverage synthetic data to overcome the limitations imposed by scarce labeled medical imaging data and enhance CellViT's segmentation capabilities.

2. Related Work

2.1 Background

Medical image segmentation plays a pivotal role in biomedical research and clinical diagnostics by enabling the accurate delineation of cellular structures and pathological regions. It supports disease detection, histopathological evaluation, and treatment planning [9], [10]. Traditional approaches, including thresholding and watershed algorithms, offered acceptable results in controlled settings but struggled with complex, real-world histological data due to overlapping structures and staining variability.

The advent of deep learning revolutionized segmentation. Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) learn hierarchical and spatial features, providing superior generalization across imaging modalities [11]. Yet, these models remain constrained by limited annotated datasets, which are costly to produce and often restricted by privacy regulations.

2.2 Synthetic Data Generation for Medical Imaging

To mitigate data scarcity, synthetic image generation has become an active area of research. Generative Adversarial Networks (GANs) and Diffusion Models (DMs) are particularly promising. While GANs synthesize plausible images, Diffusion Models outperform them in fidelity and stability [6], [13]. DiffInfinite, a hierarchical diffusion model, is designed to generate high-resolution histopathological images with corresponding masks, reducing reliance on manual annotation [7]. These synthetic images augment real datasets and support training robust segmentation models, though concerns about data memorization and generalizability persist. Recent studies apply information-theoretic metrics to evaluate the authenticity and diversity of generated images [8].

2.3 CellViT for Histopathological Image Segmentation

CellViT, a transformer-based segmentation model, leverages self-attention to capture long-range dependencies in cellular structures [3]. While it performs well across multiple datasets, its accuracy is hindered by insufficiently labeled data. Integrating synthetic data from models like DiffInfinite can improve training diversity and generalization, addressing one of the key limitations in histological segmentation.

3. Methodology

Our approach involves fine-tuning the pre-trained CellViT foundation model using different strategies, including the integration of synthetic data.

3.1. Baseline Fine-Tuning (Basic Method)

The basic fine-tuning method utilizes a labeled dataset:

$$D=\{(Ii,Mi)\}i=1N \quad (1)$$

Where Ii is a raw H&E-stained cell image and Mi is its corresponding instance-level cell mask.

1. Initialize CellViT with its pre-trained weights.
2. Optimize the model parameters using a combined Dice loss (LDice) for segmentation accuracy and Cross-entropy loss (LCE) for classification consistency. The loss function is

$$L=\lambda 1LDice+\lambda 2LCE \quad (2)$$

Where $\lambda 1$ and $\lambda 2$ are weighting coefficients.

3. Update the model weights using backpropagation until convergence. This process yields a CellViT model fine-tuned specifically for dataset D.

3.2. Fine-Tuning with Synthetic Data (Advanced Method)

This method extends the basic fine-tuning by incorporating synthetic data generated by the DiffInfinite model.

1. **Synthetic Data Generation**: Use DiffInfinite to generate synthetic H&E images Ij′.

2. **Pseudo-Labeling**: Use the baseline CellViT model to generate initial segmentation masks Mj′ for the synthetic images:

$$Ij' \ (Mj'=CellViT(Ij')) \quad (3)$$

3. **Human-in-the-Loop Verification**: Human experts review the generated pairs (Ij′,Mj′) and select high-quality samples for augmentation.
4. **Augmented Dataset Construction**: Combine the original labeled dataset D with the verified synthetic pairs to create an expanded dataset:

$$D'=D\cup\{(Ij',Mj')\}j=1M \quad (4)$$

5. **Fine-Tuning on Augmented Data**: Fine-tune CellViT on the augmented dataset D′ using the same loss function and optimization process as the basic method.

Key technical considerations for the advanced method include constructing training batches with a specific probability ratio for sampling real versus synthetic images to prevent overfitting to synthetic data, using a relatively small learning rate for stable convergence, and maximizing batch size within GPU memory limits. This approach aims to leverage the increased diversity and size of the training data to further improve segmentation performance and reliability. The overall pipeline comparing no fine-tuning, basic fine-tuning, and advanced fine-tuning is visualized below in Figure 1.

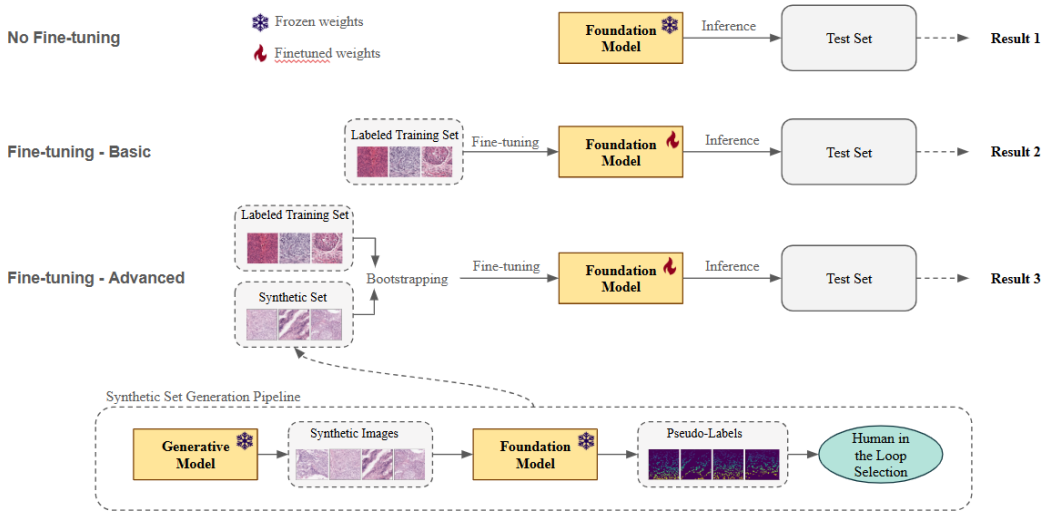

Figure 1: Research Pipeline

4. Experiments and Results

Our experimental setup utilized the MoNuSeg dataset for both training and testing, which included 30 labeled training images and 14 high-resolution test images. To augment the training data, we incorporated 28 synthetic images generated by DiffInfinite, with pseudo-segmentation labels provided by the baseline CellViT model. We evaluated performance using binary Panoptic Quality (bPQ), which combines detection and segmentation quality, and Recall, calculated as the ratio of detected true positives to all positive samples, emphasizing the importance of identifying all relevant cells. The core models employed were CellViT as the foundation model for segmentation and DiffInfinite as the generative model for synthetic data creation. Fine-tuning was conducted over 30 epochs with a learning rate of 1e-6 and a batch size of 8, using a random sampling probability of 0.6 to balance the contribution of real MoNuSeg data and synthetic DiffInfinite data during training.

4.1 Technical Considerations and Setups

Loss Function Adjustment for Fine-Tuning

During the fine-tuning process, the loss function was adjusted to accommodate the differences between the PanNuke and MoNuSeg datasets. The PanNuke dataset includes both tissue type and cell type annotations, which the original CellViT training pipeline is designed to handle. However, the MoNuSeg dataset lacks these categorical distinctions, necessitating modifications to the training pipeline to eliminate the dependency on these classifications. The loss function equation for the model is:

$$L_{total}=L_{NP}+L_{HV}+L_{NT}+L_{TC}. \quad (5)$$

We analyzed the NT (nuclei type) and TC (tissue class) loss variables, which contribute to the overall loss function for nuclei classification and tissue type identification. After identifying the corresponding lambda values, we set them to zero, effectively removing the influence of tissue and cell-type classification and ensuring the model focused on segmentation tasks relevant to MoNuSeg.

Baseline Establishment

The MoNuSeg dataset was chosen as the benchmark for evaluating segmentation performance. Initial performance was assessed using the MoNuSeg test set, providing a baseline against which improvements could be measured. Since CellViT was originally trained on the PanNuke dataset, this baseline allowed for a comparison of performance across different fine-tuning strategies.

Fine-Tuning with MoNuSeg Training Set

All MoNuSeg images were resized, and their segmentation masks were reformatted before fine-tuning. The CellViT model was fine-tuned using the MoNuSeg training set, and performance was re-evaluated on the MoNuSeg test set.

Fine-Tuning with MoNuSeg + DiffInfinite Generated Dataset

We use the DiffInfinite model to generate synthetic images, and the data is generated by solving a diffusion-based equation. The general equation for the diffusion process in generative models is:

Image Generation and Augmentation

To overcome the limitations of available training data, we utilized DiffInfinite to generate 400 synthetic histopathological images at a resolution of 512x512 pixels with 20x magnification. The MoNuSeg images were resized to match this resolution, ensuring consistency between synthetic and real datasets for model training.

Human-in-the-Loop Process

We manually reviewed and selected the DiffInfinite-generated images that most closely resembled MoNuSeg images, prioritizing those where the model captured a higher proportion of nuclei. These selected images were then segmented by CellViT, generating high-quality, pseudo-segmented images that augmented the training dataset.

Fine-Tuning with Augmented Dataset

The best pseudo-segmented images were integrated with the MoNuSeg dataset to create a hybrid MoNuSeg + DiffInfinite dataset. This augmented dataset was used for fine-tuning CellViT, improving segmentation accuracy by exposing the model to both real and synthetic data.

Random Sampling (Bootstrapped Augmentation)

To prevent overfitting to synthetic data, bootstrapped augmentation was implemented via random sampling. This ensured a balanced ratio of real and synthetic data during training, preserving generalizability while leveraging synthetic diversity.

Resizing and Preprocessing

All images in the MoNuSeg dataset were resized to 512x512 pixels to match the DiffInfinite-generated images. Additionally, MoNuSeg's segmentation masks were reformatted for compatibility with CellViT.

Loss Function Adjustment for Fine-Tuning

We made further adjustments to the loss function to accommodate MoNuSeg's lack of tissue and cell-type annotations, ensuring the model focused on segmentation tasks without relying on these classifications.

4.2 Results & Analysis

As illustrated in Figures 2 and 3, we observe a noticeable improvement in both Panoptic Quality (bPQ) and recall scores

when we fine-tune the foundation model using MonuSeg samples. These measures were mainly picked as they ensured that the model is more confident in its segmentations, allowing it to segment more cells. Through this, even with more false positives segmented, the model can catch more cells that it may have missed before. This enhancement demonstrates the benefit of domain-specific adaptation. Furthermore, we see an additional boost in performance after incorporating synthetic samples generated from DiffInfinite, which likely provides additional structural diversity and improved generalization. Overall, our results indicate a total improvement of 4% in PQ and a significant 6.9% increase in recall, highlighting the effectiveness of fine-tuning with both real and synthetic data.

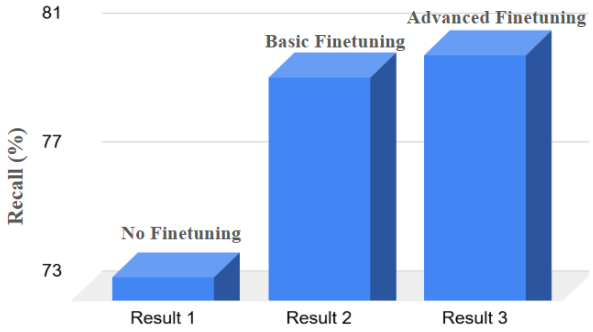

Figure 2: Quantitative Recall and bPQ scores #1

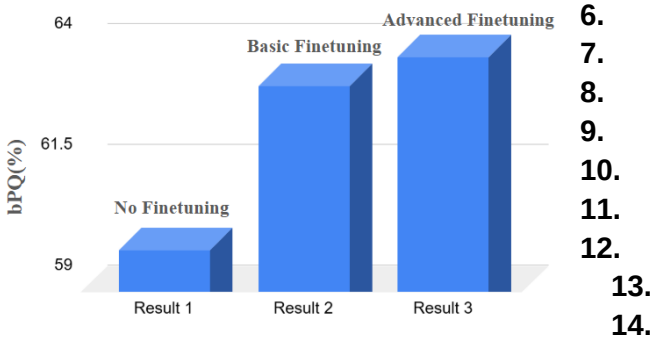

Figure 3: Quantitative Recall and bPQ scores #2

Figure 4 presents qualitative segmentation results from our test set, showcasing the model's performance in real-world scenarios. Upon visual inspection, we observe a high degree of overlap between our predicted segmentation and the corresponding ground truth. This suggests that our model can accurately delineate object boundaries, maintaining fidelity to the ground truth labels. The strong qualitative agreement further supports the quantitative gains observed in segmentation performance.

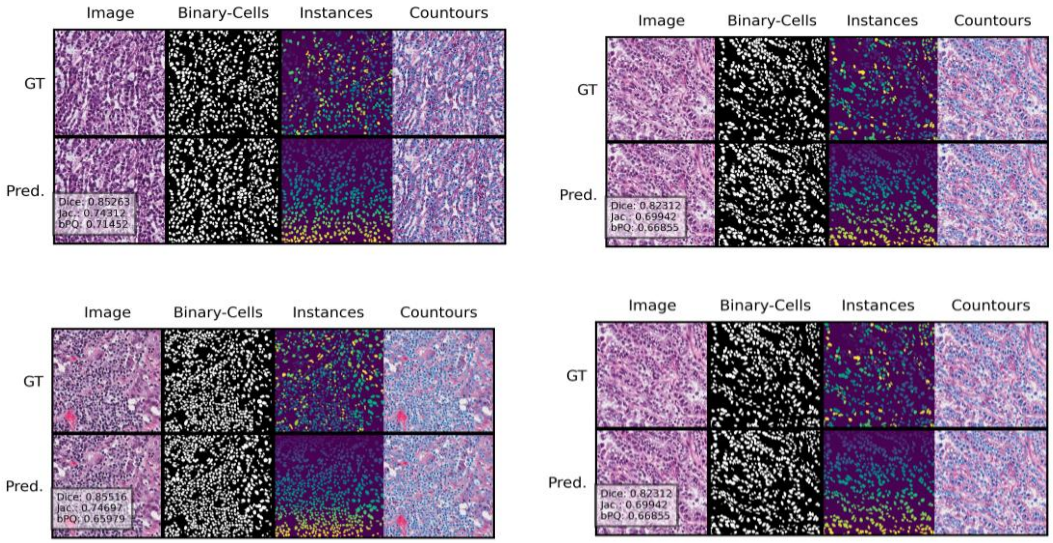

Figure 4: Qualitative segmentation results on 4 samples using our proposed method

As depicted in Figure 5, we examine the effect of altering the uniform sampling probability between synthetic and real data. Our findings reveal that as the probability of selecting real data decreases, the model becomes increasingly reliant on synthetic samples, which negatively impacts overall performance. Conversely, when the probability of selecting real data is increased, fewer synthetic samples contribute to the training process, thereby diminishing their beneficial effect. Through

systematic experimentation, we identify an optimal sampling probability of 0.6, which strikes the best balance between leveraging synthetic diversity and preserving real-data fidelity. This effect is largely due to the distillation process, where the same model is used as both the student and the teacher. Since the model learns from its own synthetic outputs, excessive reliance on synthetic data leads to compounding errors, while insufficient synthetic sampling limits the benefits of augmented diversity.

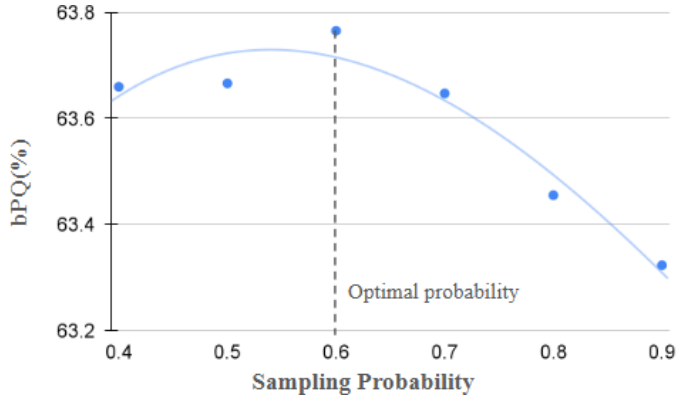

Figure 5: Varying the random sampling probability

Figure 6 illustrates the impact of varying batch sizes on model performance. We experimented with different batch sizes and observed that a batch size of 8 yielded the most optimal results. Smaller batch sizes resulted in noisier gradient updates, leading to unstable training, whereas larger batch sizes appeared to reduce the model's ability to generalize well. The batch size of 8 provided the best trade-off between stability and generalization, ensuring efficient learning.

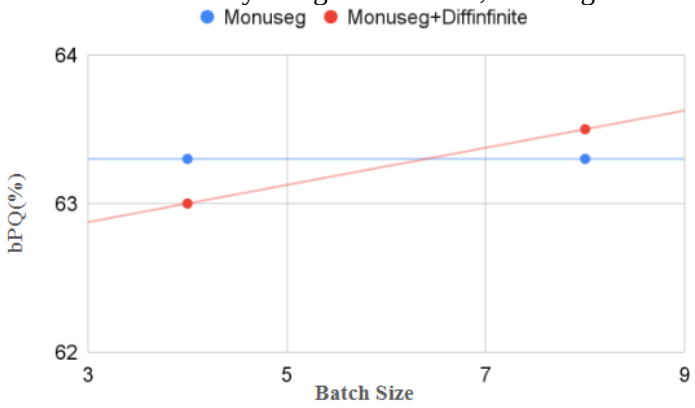

Figure 6: Varying the batch size

As shown in Figure 7, we explore the effect of adjusting the learning rate on model performance. Our analysis indicates that different learning rates lead to varying degrees of convergence efficiency and final performance. A learning rate that is too high results in unstable updates, while a rate that is too low leads to slow convergence. Through systematic evaluation, we determine that a learning rate of 5e-6 produces the best results, facilitating stable and efficient learning while preventing overfitting or underfitting.

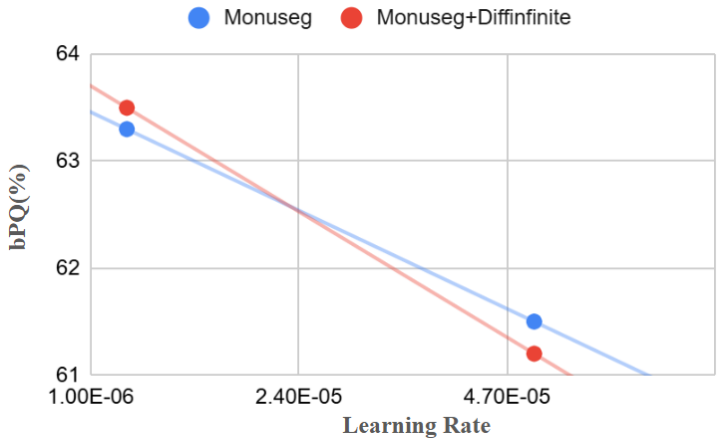

Figure 7: Varying the learning rate

5. Conclusion

Our study confirms that synthetic data generated by DiffInfinite, significantly enhances the performance of the CellViT segmentation model. The integration of synthetic data addresses the challenge of limited medical imaging datasets and demonstrates the potential of generative models in improving the performance of deep learning models for medical image analysis. The findings suggest that incorporating synthetic data into training pipelines offers a scalable solution to data scarcity, paving the way for more robust AI-driven diagnostic tools.

Acknowledgement

I would like to acknowledge the invaluable guidance and support of Mr. Darren Cameron, Prof. Zhuowen Tu, Dr. Shubhankar Bores, and Prof. Yizhe Zhang for their insightful advice, encouragement, and inspiration.

Reference

[1] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. W. K. Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images,” *Nat. Med.*, vol. 25, no. 8, pp. 1301-1309, 2019.

[2] Genomic Data Commons Data Portal. [Online]. Available: https://portal.gdc.cancer.gov/

[3] F. Hörst, M. Rempe, L. Heine, C. Seibold, J. Keyl, G. Baldini, S. Ugurel, J. Siveke, B. Grünwald, J. Egger, J. Kleesiek, “CellViT: Vision transformers for precise cell segmentation and classification,” *Med. Image Anal.*, vol. 94, pp. 103-143, 2024.

[4] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *Proc. ICML*, pp. 10347-10357, 2021.

[5] N. Kumar, R. Verma, S. Shah, A. Ramesh, S. R. Pati, M. J. Doyle, M. Feldman, A. Madabhushi, “A multi-organ nucleus segmentation challenge,” *IEEE Trans. Med. Imaging*, vol. 39, no. 5, pp. 1380-1391, 2020.

[6] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, Y. Shao, W. Zhang, B. Cui, M.-H. Yang, “Diffusion models: A comprehensive survey of methods and applications,” *ACM Comput. Surv.*, vol. 56, no. 4, pp. 1-39, 2023.

[7] M. Aversa, G. Nobis, M. Hägele, J. Chen, Y. Ma, A. Hatamizadeh, “DiffInfinite: Large mask-image synthesis via parallel random patch diffusion in histopathology,” in *Proc. NeurIPS*, vol. 36, 2023.

[8] Z. Tagmatova, A. Abdusalomov, Y. Sugimura, T. Hiroyasu, "Evaluating Synthetic Medical Images Using Artificial Intelligence with the GAN Algorithm," *Sensors*, vol. 23, no. 7, pp. 3440, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/7/3440.

[9] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60-88, 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 9351, pp. 234-241, 2015.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. ICLR*, 2021.
[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, “Segment anything,” *arXiv* preprint, arXiv:2304.02643, 2023.
[13] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *Proc. NeurIPS*, vol. 33, pp. 6840-6851, 2020.
[14] J. Ma, S. Kim, F. Li, M. Baharoon, R. Asakereh, H. Lyu, B. Wang, “Segment anything in medical images and videos: Benchmark and deployment,” *arXiv* preprint, arXiv:2408.03322, 2024.