Controllable Histopathology Image Synthesis with Training-free Structural Initialization and Textural Modulation

Yuheng Qiu[1], Jingyi Luo[1], Chenfei Ye[1], Ting Ma[1], and Jianfeng Cao[1⋆]

School of Biomedical Engineering, Harbin Institute of Technology (Shenzhen), Shenzhen, China
{qiuyuheng, 25b963007}@stu.hit.edu.cn
{yechenfei, tma, caojianfeng}@hit.edu.cn

Abstract. Deep learning has demonstrated remarkable success in high-throughput histopathology image analysis. However, the performance of learning-based models critically depends on the quality and size of annotation by expert pathologists—a resource-intensive and time-consuming process. To address the limitations of data scarcity and annotation burden, several methods have been proposed to synthesize paired histopathology data. Nevertheless, these frameworks typically still require annotation data, albeit in reduced quantities, to impose structural constraints during training. In this work, we present CHIS, a plug-in framework that guides the sampling trajectory of a pretrained diffusion model through two key stages: structural initialization at the start and textural modulation during generation. The initial noise state is refined by fusing the phase information from a prior mask with the amplitude of Gaussian noise in the frequency domain, yielding a structurally-informed starting point. During the reverse diffusion process, we adaptively modulate both coarse- and fine-grained textures at different wavelet decomposition levels. This enables a diffusion model pretrained solely on unlabeled images to generate outputs that align with prior structural masks while preserving the reference tissue style. We conducted extensive experiments demonstrating the superiority of CHIS in generation fidelity and its substantial benefits for downstream segmentation tasks. Code is available at https://github.com/IBIL-Code/CHIS.

Keywords: Histopathology Analysis · Guided Image Generation · Diffusion Model · Unsupervised Learning

1 Introduction

Histopathology remains the gold standard for diagnosis and grading because of its ability to distinguish benign from malignant tissue [17]. Learning-based algorithms now enable high-throughput analysis of whole-slide images, supporting automatic annotation and AI-assisted reporting with user-defined quantitative metrics that reduce pathologists' workload [5,10]. However, model accuracy

⋆ Corresponding author: caojianfeng@hit.edu.cn

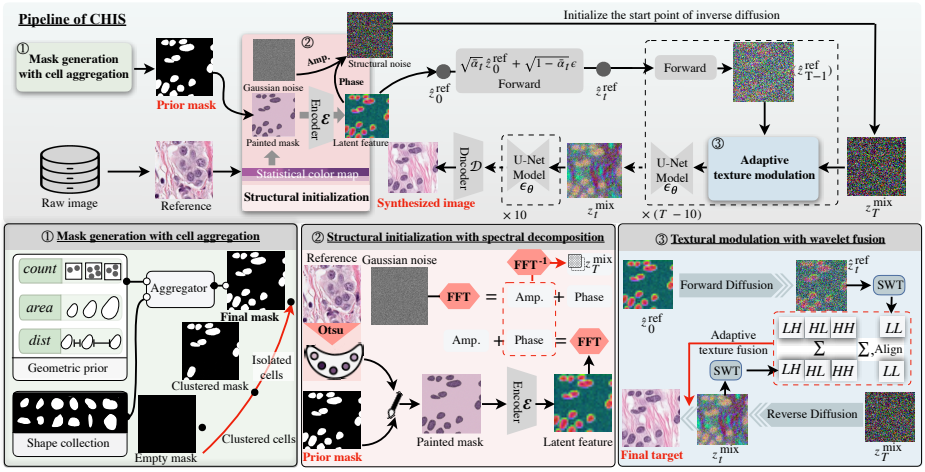

Fig. 1. Overview of our proposed CHIS for controllable histopathology image synthesis.

depends critically on the quantity and quality of labeled training data, which is expensive and time-consuming to obtain in routine clinical workflows. Synthetic data generation therefore offers a practical solution to alleviate annotation scarcity. While generative adversarial networks (GANs) can learn with weak or unpaired supervision, they often introduce artifacts that compromise diagnostic realism [3,19]. Diffusion models typically achieve higher fidelity and better conditional control, but most existing approaches require paired supervision for training [21,23]. Cycle-diffusion methods attempt unpaired translation but unfortunately suffer from noisy-clean domain mismatch, substantially increasing computational cost [27,28]. Furthermore, standard Gaussian initialization inevitably induces structural misalignment during generation [2,12].

In this work, we propose CHIS, a training-free framework that guides pretrained diffusion models to generate structure-aligned histopathology images (Fig. 1). Given a diffusion model trained solely on unlabeled histopathology images, our CHIS first generates a prior mask to regulate the spatial location of generated cells. For the starting point of the reverse sampling process, we leverage the insight that phase information primarily encodes structural contours and refine the initialization by fusing the phase of the prior mask with the magnitude of Gaussian noise in the frequency domain. Subsequently, fine-grained textures progressively interact with the dynamic sampling states through adaptive aggregation of wavelet components. The enhanced diffusion model ultimately delivers synthesized data that are not only stylistically consistent but also structurally controllable. Comprehensive evaluations demonstrate the superiority of CHIS in image quality, mask-structure fidelity, and downstream segmentation performance.

2 Method

In this section, we first introduce the pretrained diffusion model that serves as the backbone of CHIS. We then present an efficient framework for generating prior masks. In the following two subsections, we sequentially describe our strategies for refining the initialization state and guiding the reverse diffusion sampling process.

2.1 Pretrained Diffusion Model with Histopathology Images

We select the Latent Diffusion Model (LDM) [15] as our generative backbone, which shifts the learning of complex image distributions from the high-dimensional pixel space $\mathbf{X}$ to a compressed latent space $\mathbf{Z}$. Given a histopathology image $x \in \mathbf{X}$, a pretrained VAE encoder $\mathcal{E}$ maps it to a latent representation $z_0 = \mathcal{E}(x)$. The forward diffusion process progressively adds Gaussian noise to z_0 over T timesteps. For any timestep $t \in [0, T]$, the noisy latent z_t is defined as:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{1}$$

where $\bar{\alpha}_t$ follows a predefined noise schedule. The reverse denoising process follows the DDIM [18] formulation:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(z_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1}}\, \epsilon_\theta(z_t, t) \tag{2}$$

where the denoising network $\epsilon_\theta(z_t, t)$ predicts the noise added to z_t, allowing deterministic reconstruction of $\hat{z}_0$ from $z_T \sim \mathcal{N}(0, \mathbf{I})$ step-by-step. The learnable model ϵ_θ is pretrained solely on unlabeled histopathology images and remains frozen in our proposed CHIS. Finally, the VAE decoder $\mathcal{D}$ produces synthesized images by $\hat{x}_0 = \mathcal{D}(\hat{z}_0)$. In the following sections, we will describe how CHIS ensures the spatial distribution of cells in $\hat{x}_0$, starting from a random state z_T, can be precisely controlled by masks $y \in \mathbf{Y}$ generated in Section 2.2.

2.2 Generation of Histopathology Mask with Cell Aggregation

To facilitate the generation of histopathology masks, we propose an efficient framework for aggregating single-cell masks with geometric priors (① in Fig. 1). Given a collection of cell masks, the synthesis of a histopathology mask y can be formulated as an aggregation process constrained by three category-specific metrics: cell count, mask area, and inter-cell distance, denoted as $\{count, area, dist\}$. These metrics can be either derived from the mask collection or specified directly by pathologists.

To aggregate cell masks into a histopathology mask, we partition *count* cells into two groups: clustered cells and isolated cells. We first generate cell clusters by sequentially sampling cells from the collection and positioning them together within the specified *dist* threshold. Each cell cluster is then treated as a super-cell and combined with isolated cells from the collection. The distances between

all cells and super-cells are maintained above the preset *dist* threshold. Additionally, the size of each cell fluctuates around the geometric prior *area*. This generated mask benefits downstream segmentation models through densely distributed clustered cells while preserving mask fidelity through the inclusion of isolated cells.

2.3 Structural Initialization with Spectral Decomposition

During the reverse sampling process, the initial state z_T largely determines the content of the denoised results. Rather than randomly sampling z_T, we incorporate structural features from the mask y—specifically, the contours of cell masks—into z_T by decomposing y using the Fast Fourier Transform (FFT). As described in [25], structural information is primarily preserved in the phase component. We therefore refine z_T through structure-aligned phase fusion between the mask y and Gaussian noise z_T (② in Fig. 1).

For the latent diffusion model, the encoder $\mathcal{E}$ transforms the input into a latent feature map. Although $\mathcal{E}$ can directly accept the mask y, we observe a significant distribution shift in $\mathcal{E}(y)$ when a binary mask y differs substantially from the pretraining data (i.e., histopathology images). To address this issue, we apply an extract-and-fill procedure to roughly colorize the regions of y. Specifically, we first select a reference histopathology image x^{ref} and apply Otsu thresholding to obtain its coarse segmentation in the optical density domain. The background regions of y are then filled with the average color of the corresponding regions in x^{ref}. To enhance the textural features of foreground cells in y, pixels are colored according to the average color of pixels at the same boundary distance in x^{ref}. Finally, we obtain $\hat{y}$ with a color style similar to that of x^{ref}.

Based on the refined input $\hat{y}$, we obtain the latent feature $\hat{z}_0^{\text{ref}} = \mathcal{E}(\hat{y})$. The FFT operation $\mathcal{F}$ is applied to $\hat{z}_0^{\text{ref}}$ and z_T:

$$\mathcal{F}(\hat{z}_0^{\text{ref}}) = A^{\text{ref}} \exp(j\phi^{\text{ref}}), \quad \mathcal{F}(z_T) = A^{\text{noise}} \exp(j\phi^{\text{noise}}) \tag{3}$$

where $A^{\cdot}$ and $\phi^{\cdot}$ denote the amplitude and phase parts, respectively. These two parts are subsequently mixed with

$$A^{\text{mix}} = A^{\text{noise}}, \quad \phi^{\text{mix}} = \Theta(r) \odot \phi^{\text{ref}} + (1 - \Theta(r)) \odot \phi^{\text{noise}} \tag{4}$$

where $\Theta(r)$ is a low-pass filter that keeps phase only within a cutoff radius r. The structure-aligned state comes with inverse FFT $z_T^{\text{mix}} = \mathcal{F}^{-1}(A^{\text{mix}} \exp(j\phi^{\text{mix}}))$. With different r, we can control the degree of structural preservation regarding mask y. Additionally, the style of z_T^{mix} can be easily changed by selecting another x^{ref} even without any annotations.

2.4 Adaptive Textural Modulation with Wavelet Fusion

Beyond structural alignment, textural similarity also contributes to the fidelity of generated images and further reinforces structural consistency. Here we propose

wavelet-based guidance to progressively enhance textural details in z_t^{mix} and improve structure–texture coherence during the reverse diffusion process (③ in Fig. 1). We employ the stationary wavelet transform (SWT) $\mathcal{S}$ to decompose both the evolving latent z_t^{mix} and the prior guidance $\hat{z}_t^{\text{ref}}$ into coarse- and fine-grained texture components, yielding

$$\mathcal{S}(z_t^{\text{mix}}) = \{L_t^{\text{mix}}, H_t^{\text{mix}}\}, \quad \mathcal{S}(\hat{z}_t^{\text{ref}}) = \{L_t^{\text{ref}}, H_t^{\text{ref}}\} \tag{5}$$

where the coarse texture $L_t^{\cdot} = [LL_t^{\cdot}]$ and fine-grained texture $H_t^{\cdot} = [LH_t^{\cdot} \mid HL_t^{\cdot} \mid HH_t^{\cdot}]$ correspond to low- and high-frequency components in the frequency domain, respectively [8].

Since the coarse texture L_t^{mix} depicts the overall contours in z_t^{mix}, we aim to align it with L_t^{ref}. Hence, the distribution of L_t^{mix} is shifted using the mean $\mu(\cdot)$ and standard deviation $\sigma(\cdot)$ via:

$$\tilde{L}_t^{\text{mix}} = \sigma(L_t^{\text{mix}}) \, \frac{L_t^{\text{ref}} - \mu(L_t^{\text{ref}})}{\sigma(L_t^{\text{ref}})} + \mu(L_t^{\text{mix}}), \tag{6}$$

Then we modulate foreground and background textures based on three rules: i) Relaxed constraints on background textures; ii) Primary reliance on $\hat{z}_t^{\text{ref}}$ for low-frequency contour textures; iii) Reduced guidance on detailed high-frequency textures. Three hyperparameters γ, χ and ζ were set accordingly to control the degree of texture modulation. Thereafter, the final weights for modulating low- and high-frequency components are given by $\omega_L = \delta_{\{y,1\}}\chi + \delta_{\{y,0\}}\gamma$ and $\omega_H = \delta_{\{y,1\}}\zeta + \delta_{\{y,0\}}\gamma$, where δ represents the Kronecker delta function. These two weights guide the modulation with weighted summation

$$\tilde{L}_t^{\text{mix}} = \omega_L \cdot \hat{L}_t^{\text{mix}} + (1 - \omega_L) \cdot L_t^{\text{mix}}, \quad \tilde{H}_t^{\text{mix}} = \omega_H \cdot H_t^{\text{ref}} + (1 - \omega_H) \cdot H_t^{\text{mix}}, \tag{7}$$

Finally, the inverse SWT transforms the components $\tilde{L}_t^{\text{mix}}$ and $\tilde{H}_t^{\text{mix}}$ back into the spatial domain z_t^{mix}, status for the next sampling step. This process repeats during the early sampling stage, until the textures are well consolidated by $\hat{z}_t^{\text{ref}}$.

3 Experiments

3.1 Implementation Details

While CHIS is framework-agnostic across diffusion-based models, we instantiate it on PixCell [22], a backbone pretrained on large-scale histopathology images that synthesizes realistic tissue but leaves cellular structure uncontrolled. We keep PixCell *frozen*, preserving all parameters and its foundation-feature conditioning; CHIS acts only on the sampling trajectory: it replaces the initial noise at $t{=}T$ with a phase-fused latent, applies textural guidance over the first 40 steps with $\gamma = 0.2$, $\chi = 0.95$, and $\zeta = 0.6$, and lets PixCell run freely in the final 10 steps.

We evaluate CHIS on three histopathology datasets: MoNuSAC [20], Kumar [11], and PanNuke [6], with an 80/20 split for training/test data. For comparison, we select various baseline methods representing different approaches,

Table 1. Evaluation on the consistency between prior mask and synthesized image. (Best result in bold, second-best underlined. Fully supervised models are listed for reference and not for ranking.)

Method	Training	Paired data	PanNuke		MoNuSAC		Kumar	
			FS1↑	HD95↓	FS1↑	HD95↓	FS1↑	HD95↓
SDM [21]	○	○	0.7462	30.8448	0.8167	24.9139	0.7669	19.4987
NuDiff [23]	○	○	**0.8232**	**22.3510**	0.8264	22.7506	**0.7953**	**17.1287**
SynDiff [13]	○	–	0.0586	90.0501	0.0025	243.4370	0.2150	37.6211
CycleDiff [28]	○	–	0.0874	62.3175	0.2002	80.4365	0.1343	45.1336
UGDM [1]	–	–	0.1346	65.8139	0.1023	85.4937	0.1744	51.5747
ADMMDiff [26]	–	–	<u>0.5586</u>	<u>43.9161</u>	<u>0.4932</u>	<u>63.2756</u>	<u>0.5755</u>	<u>32.2030</u>
CHIS (ours)	–	–	**0.8032**	**22.6473**	**0.8671**	**14.9272**	**0.7853**	**20.3174**

including fully supervised models (NuDiff [23], SDM [21]), GAN-based models (SynDiff [13], CycleDiff [28]), and training-free models (UGDM [1], ADMMDiff [26]). For all methods, the number of synthesized images matches that of the training split in each dataset. The synthesized images are assessed from three perspectives: (i) evaluate their alignment to prior masks using a binary segmentation model; (ii) investigate their benefits to downstream instance segmentation model; (iii) measure their fidelity compared to real datasets. Detailed experimental settings are introduced in the corresponding sections.

3.2 Alignment Between Prior Mask and Synthesized Image

Inspired by [2], we adapt zero-shot nnU-Net [9] to segment the synthesized images and then evaluate the consistency between the segmentation results and the prior masks. As shown in Table 1, our CHIS achieves the best performance in preserving the structural information from the guidance masks. Although GAN-based models excel at style transformation with unpaired data, both SynDiff and CycleDiff fail to regulate mask-to-image correspondence. To refine structural details, UGDM and ADMMDiff steer the sampling trajectory by anchoring the reverse diffusion states. However, their inability to effectively modulate coarse- and fine-grained structures leads to performance degradation. Our CHIS carefully initializes the starting point with structural information and progressively modulates the dynamic states at different frequency components. The performance of our CHIS approaches that of the fully supervised model NuDiff and even exceeds it on the MoNuSAC dataset.

3.3 Improvement on Downstream Segmentation Task

To demonstrate the benefits of CHIS for downstream segmentation tasks, we augment the training dataset with synthesized images, doubling the size of each dataset. The performance of Hover-Net [7] is then compared before and after

Table 2. Benefits of synthesized images for boosting segmentation performance. (Best result in bold, second-best underlined. Fully supervised models are listed for reference and not for ranking.)

Method	Training	Paired data	PanNuke		Kumar		MoNuSAC	
			Dice↑	AJI↑	Dice↑	AJI↑	Dice↑	AJI↑
Baseline			0.7742	0.5822	0.8078	0.5746	0.7741	0.5861
CutOut [4]	–	–	0.8160	0.6343	0.8244	0.6063	<u>0.7883</u>	<u>0.6046</u>
CutMix [24]	–	–	<u>0.8189</u>	<u>0.6368</u>	<u>0.8256</u>	<u>0.6089</u>	0.7885	0.6021
SDM [21]	○	○	0.8128	0.6375	**0.8280**	**0.6252**	**0.7945**	0.6074
NuDiff [23]	○	○	0.8223	0.6422	0.8242	0.6097	0.7930	**0.6101**
CycleDiff [28]	○	–	0.8116	0.6259	0.8191	0.6015	0.7766	0.5896
SynDiff [13]	○	–	0.8186	0.6348	0.8228	0.6044	0.7883	0.6012
UGDM [1]	–	–	0.8152	0.6324	0.8212	0.6031	0.7806	0.5884
ADMMDiff [26]	–	–	0.8181	0.6366	0.8213	0.6082	0.7873	0.6025
CHIS (Ours)	–	–	**0.8224**	**0.6437**	**0.8271**	**0.6238**	**0.7895**	**0.6046**

Table 3. Image quality assessment of synthesized data. (Best result in bold, second-best underlined. Fully supervised models are listed for reference and not for ranking.)

Method	Training	Paired data	PanNuke		MoNuSAC		Kumar	
			FID↓	IS↑	FID↓	IS↑	FID↓	IS↑
SDM [21]	○	○	10.0452	2.6973	5.7759	**3.2963**	7.8278	2.9161
NuDiff [23]	○	○	10.6155	3.4152	7.5344	2.9655	6.3651	2.6072
CycleDiff [28]	○	–	12.0908	2.5262	12.2660	2.0854	12.2660	2.3193
SynDiff [13]	○	–	13.0348	1.5896	11.1433	1.3745	11.1433	1.7235
UGDM [1]	–	–	<u>7.7972</u>	3.0428	5.9682	2.5637	<u>5.9761</u>	2.3408
ADMMDiff [26]	–	–	9.1775	<u>3.1139</u>	<u>5.4475</u>	<u>3.1627</u>	6.3033	<u>2.8059</u>
CHIS (Ours)	–	–	**7.7772**	**4.2514**	**5.4099**	**3.2250**	**5.8926**	**3.4731**

dataset expansion. In addition to the methods mentioned in Section 3.2, we also include comparisons with on-the-fly augmentation techniques CutOut [4] and CutMix [24]. As shown in Table 2, CHIS achieves the highest Dice and AJI scores among all competitive methods. Compared to the baseline, the segmentation performance gain from our framework is substantial, especially considering that it eliminates the need for data annotation and model training. Moreover, we observe that CHIS's performance approaches that of fully supervised models SDM and NuDiff, which require substantial annotation and computational resources. The performance difference of NuDiff between Table 1 and Table 2 indicates that synthesized images require not only overall contour alignment but also similarity in textural features. Only by achieving both aspects can we avoid

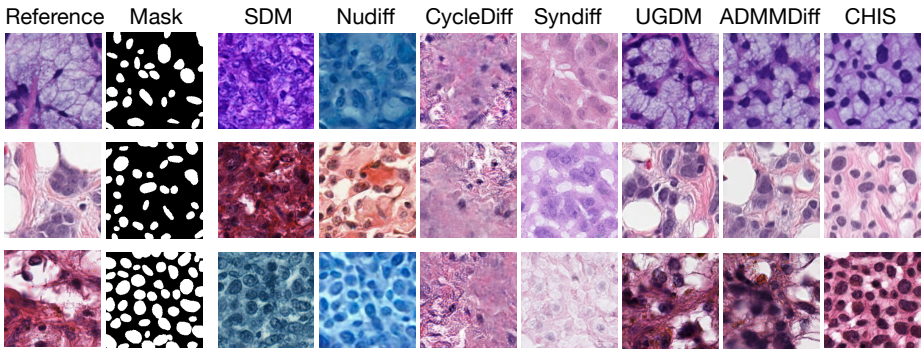

Fig. 2. Comparison of synthesized images from different methods.

introducing distribution shifts during model testing. This finding indirectly validates CHIS's ability to balance both global structural integration and local feature preservation.

3.4 Image Quality Assessment of Synthesized Data

In this section, we compare the synthesized data directly with the real image dataset, where similarity is quantified using the image quality metrics FID computed with CLIP[14] and IS [16]. When a style reference image is required for UGDM, ADMMDiff, and CHIS, we randomly select one from the corresponding training set. Quantitative results in Table 3 demonstrate the advantage of CHIS in generating high-quality images with respect to visual perception. The improved performance of UGDM, ADMMDiff, and CHIS shows that incorporating a reference during the reverse sampling process effectively regulates the style of generated content. In contrast, both supervised and GAN-based models focus on learning general distribution features from heterogeneous data, which inevitably leads to synthetic data exhibiting artifacts with mixed styles. In Fig. 2, we present three example results with their corresponding style references and prior masks. Our CHIS achieves superior performance in terms of both cell distribution and style preservation.

4 Conclusion

In this work, we explore the feasibility of a controllable framework for histopathology image synthesis, addressing the critical challenge of limited data resources in clinical applications. Equipped with our training-free structural initialization and textural modulation modules, pretrained diffusion models can now achieve high mask-to-image faithfulness without requiring labor-intensive data annotation or substantial computational overhead. Experimental results demonstrate that our proposed CHIS not only rivals competitive works in generation quality, but also impressively improves downstream instance segmentation performance.

As an efficient and scalable solution that bridges the gap between zero-shot efficiency and domain-specific precision, CHIS has the potential to advance various downstream applications, such as virtual staining and pathologist training.

Acknowledgments. This work was funded by Guangdong Basic and Applied Basic Research Foun- dation (No. 2026A1515010255) and Shenzhen Science and Technology Program (No. JCYJ20250604145623032).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 843–852 (2023)
2. Bhosale, M., Wasi, A., Zhai, Y., Tian, Y., Border, S., Xi, N., Sarder, P., Yuan, J., Doermann, D., Gong, X.: Pathdiff: Histopathology image synthesis with unpaired text and mask conditions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22415–22424 (2025)
3. Butte, S., Wang, H., Xian, M., Vakanski, A.: Sharp-gan: Sharpness loss regularized gan for histopathology image synthesis. In: 2022 IEEE 19th International symposium on biomedical imaging (ISBI). pp. 1–5. IEEE (2022)
4. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017), `https://arxiv.org/abs/1708.04552`
5. Ding, T., Wagner, S.J., Song, A.H., Chen, R.J., Lu, M.Y., Zhang, A., Vaidya, A.J., Jaume, G., Shaban, M., Kim, A., et al.: A multimodal whole-slide foundation model for pathology. Nature medicine pp. 1–13 (2025)
6. Gamper, J., Koohbanani, N.A., Benet, K., Khuram, A., Rajpoot, N.: Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: European Congress on Digital Pathology. pp. 11–19. Springer (2019)
7. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N.: Hover-net: Simultaneous segmentation and classification of nuclei in multitissue histology images. Medical image analysis **58**, 101563 (2019)
8. Huang, W.C., Chang, L.W.: Predictive subband image coding with wavelet transform. Signal processing: Image communication **13**(3), 171–181 (1998)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Janowczyk, A., Madabhushi, A.: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics **7**(1), 29 (2016)
11. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE transactions on medical imaging **36**(7), 1550–1560 (2017)
12. Li, Y., Shao, H.C., Liang, X., Chen, L., Li, R., Jiang, S., Wang, J., Zhang, Y.: Zero-shot medical image translation via frequency-guided diffusion models. IEEE transactions on medical imaging **43**(3), 980–993 (2023)

13. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Cukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
15. 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)
16. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
17. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
18. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022), https://arxiv.org/abs/2010.02502
19. Tschuchnig, M.E., Oostingh, G.J., Gadermayr, M.: Generative adversarial networks in digital pathology: a survey on trends and future potential. Patterns **1**(6) (2020)
20. Verma, R., Kumar, N., Patil, A., Kurian, N.C., Rane, S., Graham, S., Vu, Q.D., Zwager, M., Raza, S.E.A., Rajpoot, N., Wu, X., Chen, H., Huang, Y., Wang, L., Jung, H., Brown, G.T., Liu, Y., Liu, S., Jahromi, S.A.F., Khani, A.A., Montahaei, E., Baghshah, M.S., Behroozi, H., Semkin, P., Rassadin, A., Dutande, P., Lodaya, R., Baid, U., Baheti, B., Talbar, S., Mahbod, A., Ecker, R., Ellinger, I., Luo, Z., Dong, B., Xu, Z., Yao, Y., Lv, S., Feng, M., Xu, K., Zunair, H., Hamza, A.B., Smiley, S., Yin, T.K., Fang, Q.R., Srivastava, S., Mahapatra, D., Trnavska, L., Zhang, H., Narayanan, P.L., Law, J., Yuan, Y., Tejomay, A., Mitkari, A., Koka, D., Ramachandra, V., Kini, L., Sethi, A.: Monusac2020: A multi-organ nuclei segmentation and classification challenge. IEEE Transactions on Medical Imaging **40**(12), 3413–3423 (2021). https://doi.org/10.1109/TMI.2021.3085712
21. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic image synthesis via diffusion models. arXiv preprint arXiv:2207.00050 (2022)
22. Yellapragada, S., Graikos, A., Li, Z., Triaridis, K., Belagali, V., Kapse, S., Nandi, T.N., Madduri, R.K., Prasanna, P., Kurc, T., et al.: Pixcell: A generative foundation model for digital histopathology images. arXiv preprint arXiv:2506.05127 (2025)
23. Yu, X., Li, G., Lou, W., Liu, S., Wan, X., Chen, Y., Li, H.: Diffusion-based data augmentation for nuclei image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 592–602. Springer (2023)
24. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 6023–6032 (2019). https://doi.org/10.1109/ICCV.2019.00612
25. Zeng, Y., Ochoa, C., Zhou, M., Patel, V.M., Guizilini, V., McAllister, R.: Neuralremaster: Phase-preserving diffusion for structure-aligned generation. arXiv preprint arXiv:2512.05106 (2025)
26. Zhang, Y., Liu, Z., Li, Z., Li, Z., Clark, J.J., Si, X.: Decoupling training-free guided diffusion by admm. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23292–23302 (2025)

27. Zhu, P., Liu, C., Fu, Y., Chen, N., Qiu, A.: Cycle-conditional diffusion model for noise correction of diffusion-weighted images using unpaired data. Medical image analysis p. 103579 (2025)
28. Zou, S., Huang, Y., Yi, R., Zhu, C., Xu, K.: Cyclediff: Cycle diffusion models for unpaired image-to-image translation. arXiv preprint arXiv:2508.06625 (2025)