CASCADED DIFFUSION MODELS FOR 2D AND 3D MICROSCOPY IMAGE SYNTHESIS TO ENHANCE CELL SEGMENTATION

Rüveyda Yilmaz (✉), Kaan Keven, Yuli Wu, Johannes Stegmaier

Institute of Imaging and Computer Vision, RWTH Aachen University, Germany
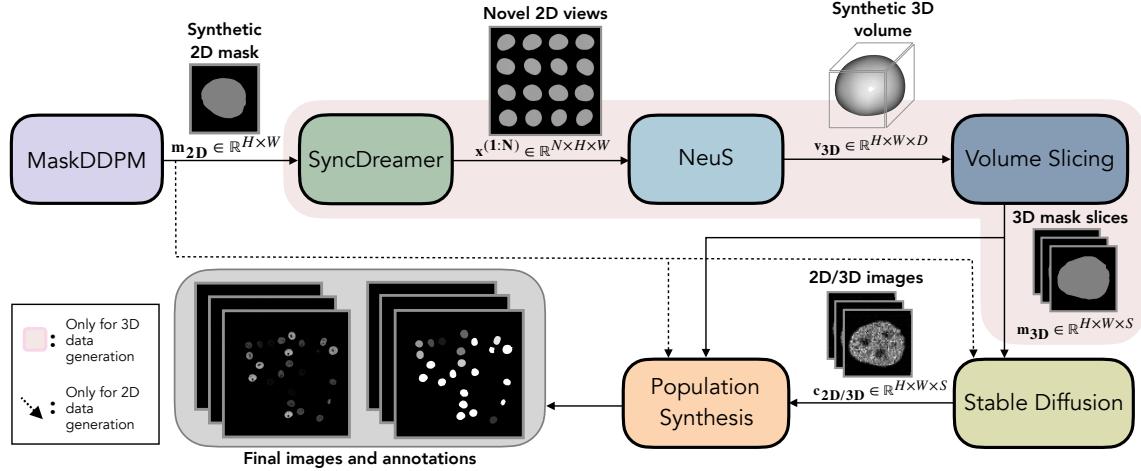Email: *rueveyda.yilmaz@lfb.rwth-aachen.de*

ABSTRACT

Automated cell segmentation in microscopy images is essential for biomedical research, yet conventional methods are labor-intensive and prone to error. While deep learning-based approaches have proven effective, they often require large annotated datasets, which are scarce due to the challenges of manual annotation. To overcome this, we propose a novel framework for synthesizing densely annotated 2D and 3D cell microscopy images using cascaded diffusion models. Our method synthesizes 2D and 3D cell masks from sparse 2D annotations using multi-level diffusion models and NeuS, a 3D surface reconstruction approach. Following that, a pretrained 2D Stable Diffusion model is finetuned to generate realistic cell textures and the final outputs are combined to form cell populations. We show that training a segmentation model with a combination of our synthetic data and real data improves cell segmentation performance by up to 9% across multiple datasets. Additionally, the FID scores indicate that the synthetic data closely resembles real data. The code for our proposed approach will be available at https://github.com/ruveydayilmaz0/cascaded_diffusion.

Index Terms— Diffusion models, 2D and 3D microscopy image synthesis, Cell segmentation

1. INTRODUCTION

Automatic cell segmentation in microscopy images enables the quantitative analysis of cellular structures and processes, advancing biomedical research in domains such as cancer diagnosis [1], drug discovery [2], and pathological assessment [3]. Conventional methods, such as manual annotation or classical image processing techniques, are time-consuming, prone to human error, and often struggle with the complexity and variability of biological data [4]. Deep learning-based methods, on the other hand, have demonstrated the ability to automatically extract detailed features from images and accurately delineate individual cells, even in complex and noisy environments [3, 2, 1]. Despite their success, they generally require large annotated microscopy data for training which is usually scarce due to the difficult manual annotation. One approach to solve this problem is to use synthetically generated data, which can supplement real datasets and provide diverse, labeled examples for training, improving model generalization and performance [4]. To address this, [5] introduces a synthetic image generation method for fluorescence microscopy images. The authors assume convex shapes for real cells and generate random 3D ellipsoids to represent cell volumes, which are subsequently used as annotation labels for the synthetic dataset. To generate the final images conditioned on these randomly created volumes, they propose an architecture called SpCycleGAN, an extension of CycleGAN [6]. Chen *et al.* criticize the use of ellipsoidal cell nuclei volumes, arguing that this approach is inadequate for representing nonconvex nuclei [7]. As an alternative, they propose employing Bézier curves to create synthetic nuclei shapes. Constrained by these shapes, they generate the final volumes using the SpCycleGAN architecture [5]. Similarly, [8] adopts an approach akin to [5] for simulating nuclei shapes, but utilizes denoising diffusion probabilistic models (DDPMs) [9] to overlay texture onto the generated shapes. In addition to generating 3D cell microscopy images, the proposed model also synthesizes 2D videos of living cells modifying the UNet architecture from the DDPM. Alternatively, [10] presents an approach for generating 2D cell microscopy image sequences by employing a 2D DDPM alongside a flow prediction model to produce temporally consistent synthetic sequences. While the approaches mentioned above offer methods to generate synthetic cell microscopy data in 2D or 3D, they either rely on assumed cell shapes to generate masks or base the masks on real data statistics. The former approach does not incorporate actual information from real cell shapes, which can lead to unrealistic cell structures due to the imposed assumptions. The latter approach utilizes real nuclei shape information; however, many 3D datasets lack comprehensive 3D annotations, with labels typically available only for a single z-slice [11]. This limitation makes it infeasible to train models on fully-annotated 3D data to generate synthetic 3D masks. Moreover, [7, 8, 5, 10] train generative models from scratch to synthesize cell textures. Yet, the performance could potentially be enhanced by finetuning a pretrained foundation model with the limited data available [12]. In this paper, we propose an approach for synthesizing densely annotated 2D and 3D cell microscopy images with a cascade of diffusion models and volume recon-

Fig. 1: Overview of the proposed method. For 2D data synthesis, MaskDDPM (▢) and Stable Diffusion (▢) generate masks and cell textures respectively. For 3D data generation, SyncDreamer, NeuS and volume slicing (▢) are additionally employed. The final images and the masks are combined using the population synthesis module (▢) .

struction. Specifically, we generate synthetic 3D masks from sparse 2D real annotations, introduce a method to finetune the pretrained 2D Stable Diffusion model for 2D and 3D cell texture synthesis and demonstrate that segmentation performance can be improved using the synthetic data produced by the proposed methodology.

2. METHODS

Mask Synthesis in 2D. To define the cell outlines ($m_{2D} \in \mathbb{R}^{H \times W}$), we start by generating synthetic masks using a limited number of ground truth annotations from real datasets. Accordingly, we employ a DDPM architecture, which we name MaskDDPM (Fig. 1, ▢), as an initial step towards 2D and 3D mask synthesis [13]. In many 3D microscopy datasets, only a single slice in the z-dimension is sparsely annotated [11], making it impractical to train a 3D model due to the lack of complete cell shape information. Thus, we also use a 2D MaskDDPM architecture for 3D datasets by training it on the sparse 2D annotations. Although the output is always 2D, this initial step serves as a foundation for both 2D and 3D image synthesis.

Multiview-consistent Mask Generation. For generating 3D synthetic data, we utilize SyncDreamer (Fig. 1, ▢) to predict novel 2D views ($x^{(1:N)} \in \mathbb{R}^{N \times H \times W}$) from the 2D output mask m_{2D} produced by MaskDDPM [14]. SyncDreamer, conditioned on a single 2D image of an object, generates multiple unseen views simultaneously using a DDPM architecture. To ensure multiview consistency, noise across multiple views is jointly predicted according to the following DDPM training objective [14]:
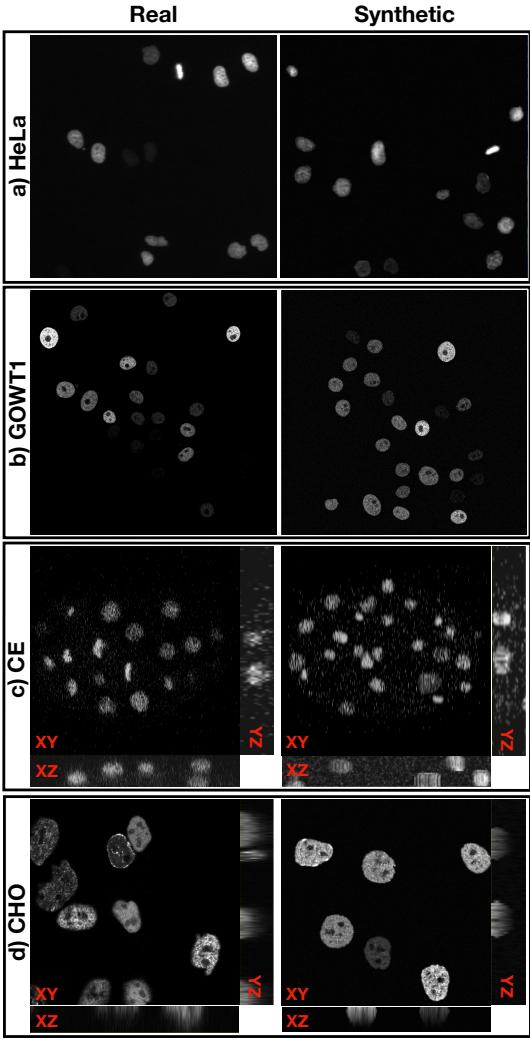
$$L(\theta) = \mathbb{E}_{t, x_0^{(1:N)}, n, \epsilon^{(1:N)}} \left[\left\| \epsilon^{(n)} - \epsilon_\theta^{(n)} \left(x_t^{(1:N)}, t \right) \right\|_2 \right], \quad (1)$$

where x_t is the noisy image at diffusion timestep t, N is the number of predicted 2D views, $\epsilon^{(n)}$ and $\epsilon_\theta^{(n)}$ are the added and the predicted noises for the view n, respectively.

The model is designed to be trained on random 2D views of 3D volumes. However, as previously mentioned, dense 3D masks are often unavailable in real microscopy sequences. To address this, we propose training SyncDreamer on spherical harmonic shapes [15]. Specifically, we generate spherical harmonic volumes and extract multiple 2D surface images from random viewing directions for each volume. These 2D images form a training dataset that enables SyncDreamer to learn how to predict additional views of a shape from a single 2D view. During inference, conditioned on m_{2D}, the output from MaskDDPM, SyncDreamer predicts multiple novel views for the corresponding hypothetical 3D volume.

Multiview-image to Volume Generation. To generate a dense volumetric mask ($v_{3D} \in \mathbb{R}^{H \times W \times D}$) from the predicted multiple views, we employ NeuS ([16], Fig. 1 ▢). NeuS is a surface reconstruction method that represents object surfaces as signed distance functions, modeled by multi-layer perceptrons (MLPs). The MLPs are trained to align the rendered surface images with the provided multiview inputs. Using the multiview 2D mask predictions $x^{(1:N)}$ from SyncDreamer, we construct 3D volumes v_{3D} via NeuS. Consequently, rather than relying on randomly generated spheres as masks for synthetic data [17, 5, 18], our approach transfers information from sparse 2D ground truth annotations into the synthetic data generation process.

Slicing the Synthetic Masks. Since real 3D cell microscopy images typically consist of multiple 2D slices representing a 3D volume along the z-dimension [11], we slice the output volumes v_{3D} from NeuS at equal intervals ($m_{3D} \in \mathbb{R}^{H \times W \times S}$, where S is the number of slices) to align their structure with that of the real datasets (Fig. 1, ▢).

Fig. 2: Representative examples from the real and synthetic (a) HeLa, (b) GOWT1, (c) CE, and (d) CHO datasets. For the 3D CE and CHO datasets, zoomed-in 2D cross-sectional images along the X and Y dimensions are also provided.

Cell Texture Generation. To apply realistic cell textures ($c_{2D/3D} \in \mathbb{R}^{H \times W \times S}$) to the shapes defined by the synthetic masks m_{2D} or m_{3D}, we utilize Stable Diffusion, a foundational text-to-image diffusion model [19]. For microscopy image generation, it is crucial that the cell shapes in the generated images accurately correspond to the synthetic masks, as these image-mask pairs are usually intended for training segmentation models. To ensure this alignment, we employ a modified version of the Stable Diffusion architecture ([12], Fig. 1 □). Specifically, instead of using text prompts, we condition the model on synthetic masks and allow it to overlay realistic cell textures. This is achieved by copying the UNet weights from the pretrained Stable Diffusion model and connecting it to the image-based conditional input via convolutional blocks at multiple scales. During training, these additional convolutional blocks are learned, while the

pretrained UNet weights from Stable Diffusion are finetuned using real cell microscopy images and their ground truth annotations as shape-conditioned inputs.

Stable Diffusion is designed for generating 2D outputs, but not for 3D data generation. To harness its capabilities for 3D image synthesis, we propose a method that enables the use of the 2D model for 3D data generation. We generate images for each slice of the 2D synthetic masks using the mask-conditional Stable Diffusion model that we finetuned. However, generating 2D slice images independently would disrupt the continuity of textures across slices for each cell. To preserve this continuity, we exploit the inherent properties of Stable Diffusion, which operates as a Denoising Diffusion Implicit Model (DDIM) [13] when the hyperparameter η, controlling the stochasticity of the diffusion process, is set to 0. In DDIMs, changes in the latent noise space are deterministically reflected in the image space, as randomness is eliminated by setting $\eta = 0$. Through this property, the latents $c_T^0, ..., c_T^S$ can be manipulated to have the desired amount of correlation in synthetic images $c_0^0, ..., c_0^S$. Correspondingly, inspired by [20], we set a common noise vector c_{cmn} that is shared among all the slices of each 3D image, and a unique noise vector c_{unq}^s that is independently generated for each slice. Combinations of these noise vectors are used to generate the latent noise samples $c_T^0, ..., c_T^S$ as follows:

$$c_{cmn}, c_{unq}^s \sim \mathcal{N}(0, I),$$

$$c_T^s = \frac{\rho}{\sqrt{1+\rho^2}} \cdot c_{cmn} + \frac{1}{\sqrt{1+\rho^2}} \cdot c_{unq}^s \quad, \qquad (2)$$

where s is the slice number and ρ is the strength of the texture consistency among slices. The noise vectors are scaled by constants dependent on ρ such that c_T^s has unit variance.

Using the approach outlined in Equation 2, the slices from each synthetic image c_{3D} exhibit structural similarity through c_{cmn}, while also incorporating some variation through c_{unq}^s.
Cell Population Synthesis. As the process of data synthesis described thus far involves one cell per image, we ultimately merge the outputs to create cell populations. Following the approach proposed in [18], we initialize two empty canvases for each synthetic raw image-mask pair ($c_{2D/3D}, m_{2D/3D}$). Random synthetic cells are iteratively placed on the raw image canvas at unoccupied locations, with the corresponding masks aligned at the same positions on the mask canvas. To simulate the clustering of cells as observed in real datasets, each new cell is positioned near the previously added cells with a certain probability. This probability is modeled using a Bernoulli distribution with parameter $p \in [0, 1]$, referred to as the *clustering probability*.

3. EXPERIMENTS

Datasets. To demonstrate the proposed methodology, we experiment with four public cell nuclei datasets made available

Table 1: Qualitative results for (a) the segmentation performance (SEG), (b) FID, and (c) ablation experiments.

Dataset	None	Real	Synthetic	Both	FID_{r2r}	FID_{r2s}	w/o SyncDr.	w/o pretr.
HeLa	0.775	0.826	0.869	**0.877**	5.7	6.1	-	0.843
GOWT1	0.471	0.881	0.905	**0.914**	0.9	3.4	-	0.889
CHO	0.698	0.897	0.893	**0.909**	0.3	0.7	0.876	0.890
CE	0.212	0.706	0.738	**0.793**	1.2	2.5	0.788	0.791

(a) The segmentation performance (SEG ↑) of Cellpose generalist model without finetuning (None), with finetuning on real, synthetic, and both datasets.

(b) The FID score (↓) between two sets of real data and between real and synthetic data.

(c) The segmentation performance (SEG ↑) of Cellpose in ablation experiments without SyncDreamer (w/o SyncDr.) or pretrained Stable Diffusion (w/o pretr.).

by the Cell Tracking Challenge [11]. Each dataset comprises two grayscale confocal microscopy image sequences, with lengths ranging from 92 to 250 frames. These datasets include 2D images of HeLa cancer cells expressing the H2B-GFP protein and mouse stem cells expressing GFP-Oct4, 3D images of Chinese hamster ovary cells expressing GFP-PCNA and developing embryo cells of *C. elegans* expressing H2B-GFP. For convenience, we abbreviate them as HeLa, GOWT1, CHO, and CE, respectively, throughout this paper. The datasets include manually annotated sparse gold truth labels and densely annotated silver truth labels, which are obtained through a voting mechanism among multiple cell segmentation models [11]. In each dataset, we use one of the annotated sequences for training and the other for testing. Before training, we identify the cells in the images using the silver truth masks and create crops of size 128×128 for both the cells and their corresponding silver truth masks. This step is essential, as SyncDreamer cannot predict multiview images when multiple masks are present in the same frame. The generated mask crops are subsequently used for training MaskDDPM, while the image-mask pairs are employed for finetuning Stable Diffusion.

Results. To evaluate the quality of our synthetic data, we experiment with a cell segmentation method called Cellpose [21]. As Cellpose is a generalist model designed to segment unseen data, we first assess its performance on one of the annotated real sequences, which we reserve solely for testing. Next, we finetune it on the remaining real sequence, on 1000 synthetic images, and on the mixture of both real and synthetic data. As a result, we have four Cellpose models trained with different combinations of data for each of the four datasets. To assess the performance of each model, we compute SEG, which calculates the intersection over union (IoU) between the predicted and the ground truth segmentation masks [11]. The results given in Table 1a indicate that Cellpose does not consistently perform well without finetuning, particularly for the GOWT1 and CE datasets. Notably, finetuning the model with either real or synthetic data significantly enhances the performance, with the best results achieved using both.

As another evaluation metric, we compute the Fréchet Inception Distance (FID) on real and synthetic data [22]. First, to establish a baseline for the typical distance between two sets of real data, we calculate the FID between the real training and the test sets and name this FID_{r2r}. Next, we calculate the FID between the real test data and 1000 synthetic data, and denote this as FID_{r2s}. As shown in Table 1b, FID_{r2r} and FID_{r2s} are sufficiently close, suggesting that the synthetic data can be considered realistic, as discussed in [23]. In Fig. 2, we present representative example images from the real and synthetic datasets.

Ablation experiments. To assess the contribution of our proposed 3D mask generation approach (Fig. 1, ▢) and the pretrained Stable Diffusion (Fig. 1, ▢) on the overall performance, we conducted ablation experiments by replacing the output from NeuS (Fig. 1, $\mathbf{v_{3D}}$) by spherical harmonic shapes and Stable Diffusion by a Latent Diffusion Model (LDM) [19] without pretraining. Subsequently, we train Cellpose with a mixture of the resulting synthetic data and the real data. As shown in Table 1c, replacing $\mathbf{v_{3D}}$ while keeping the other components unchanged leads to a reduction in segmentation performance for both the CHO and CE datasets. This ablation experiment is not applicable to 2D datasets, as the ablated components are not used for 2D image synthesis in the original pipeline. When an LDM is trained from scratch for the same number of epochs as Stable Diffusion, a comparable decline in performance is observed, and it requires significantly longer training to achieve the same performance as the finetuned Stable Diffusion.

4. CONCLUSION

In this work, we introduce a novel pipeline to synthesize 2D and 3D microscopy images using *ad hoc* diffusion models for generating single-view 2D masks with MaskDDPM, 3D mask volumes with SyncDreamer and NeuS, and cell textures with Stable Diffusion. The resulting 3D volumes are geometrically consistent due to high multiview coherence and both 2D and 3D data are realistic as evidenced by the FID score. Besides, the synthetic data is capable of augmenting real datasets, leading to improved segmentation accuracy across various cell microscopy datasets as the ultimate goal. Future work will explore extending the pipeline to different biomedical imaging modalities and applying it to time-series data.

5. COMPLIANCE WITH ETHICAL STANDARDS

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] Pin Wang, Xianling Hu, Yongming Li, Qianqian Liu, and Xinjian Zhu, "Automatic cell nuclei segmentation and classification of breast cancer histopathology images," *Signal Processing*, vol. 122, pp. 1–13, 2016.

[2] Daniel Krentzel, Spencer L Shorte, and Christophe Zimmer, "Deep learning in image-based phenotypic drug discovery," *Trends in Cell Biology*, vol. 33, no. 7, pp. 538–554, 2023.

[3] Abdulkadir Albayrak and Gokhan Bilgin, "Automatic cell segmentation in histopathological images via two-staged superpixel-based algorithms," *Medical & Biological Engineering & Computing*, vol. 57, pp. 653–665, 2019.

[4] Zhichao Liu, Luhong Jin, Jincheng Chen, Qiuyu Fang, Sergey Ablameyko, Zhaozheng Yin, and Yingke Xu, "A survey on applications of deep learning in microscopy image analysis," *Computers in Biology and Medicine*, vol. 134, pp. 104523, 2021.

[5] Chichen Fu, Soonam Lee, David Joon Ho, Shuo Han, Paul Salama, Kenneth W Dunn, and Edward J Delp, "Three dimensional fluorescence microscopy image synthesis and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2221–2229.

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[7] Alain Chen, Liming Wu, Shuo Han, Paul Salama, Kenneth W Dunn, and Edward J Delp, "Three dimensional synthetic non-ellipsoidal nuclei volume generation using bezier curves," in *IEEE 18th International Symposium on Biomedical Imaging*, 2021, pp. 961–965.

[8] Dennis Eschweiler, Rüveyda Yilmaz, Matisse Baumann, Ina Laube, Rijo Roy, Abin Jose, Daniel Brückner, and Johannes Stegmaier, "Denoising diffusion probabilistic models for generation of realistic fully-annotated microscopy image datasets," *PLOS Computational Biology*, vol. 20, no. 2, pp. e1011890, 2024.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[10] Rüveyda Yilmaz, Dennis Eschweiler, and Johannes Stegmaier, "Annotated biomedical video generation using denoising diffusion probabilistic models and flow fields," in *International Workshop on Simulation and Synthesis in Medical Imaging*. Springer, 2024, pp. 197–207.

[11] Martin Maška, Vladimír Ulman, Pablo Delgado-Rodriguez, Estibaliz Gómez-de Mariscal, Tereza Nečasová, Fidel A Guerrero Peña, Tsang Ing Ren, Elliot M Meyerowitz, Tim Scherr, Katharina Löffler, et al., "The cell tracking challenge: 10 years of objective benchmarking," *Nature Methods*, vol. 20, no. 7, pp. 1010–1020, 2023.

[12] 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*, 2023, pp. 3836–3847.

[13] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[14] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang, "Syncdreamer: Generating multiview-consistent images from a single-view image," in *The Twelfth International Conference on Learning Representations*, 2024.

[15] Claus Müller, *Spherical harmonics*, vol. 17, Springer, 2006.

[16] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang, "Neus: Learning neural implicit surfaces by volume rendering for multiview reconstruction," in *Advances in Neural Information Processing Systems*, 2021.

[17] Dennis Eschweiler, Malte Rethwisch, Mareike Jarchow, Simon Koppers, and Johannes Stegmaier, "3d fluorescence microscopy data synthesis for segmentation and benchmarking," *PLOS ONE*, vol. 16, no. 12, pp. e0260509, 2021.

[18] David Svoboda and Vladimír Ulman, "Towards a realistic distribution of cells in synthetically generated 3d cell populations," in *Image Analysis and Processing–ICIAP 2013: 17th International Conference, Naples, Italy, September 9-13, 2013, Proceedings, Part II 17*. Springer, 2013, pp. 429–438.

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10684–10695.

[20] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin

Huang, Ming-Yu Liu, and Yogesh Balaji, "Preserve your own correlation: A noise prior for video diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22930–22941.

[21] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu, "Cellpose: a generalist algorithm for cellular segmentation," *Nature Methods*, vol. 18, no. 1, pp. 100–106, 2021.

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] Prafulla Dhariwal and Alexander Nichol, "Diffusion models beat GANs on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.