This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Volume Tells: Dual Cycle-Consistent Diffusion for 3D Fluorescence Microscopy De-noising and Super-Resolution

Zelin Li[1,2,*], Chenwei Wang[2,*], Zhaoke Huang[1,2,†], Yiming Ma[3], Cunming Zhao[3], Zhongying Zhao[3], Hong Yan[1,2]

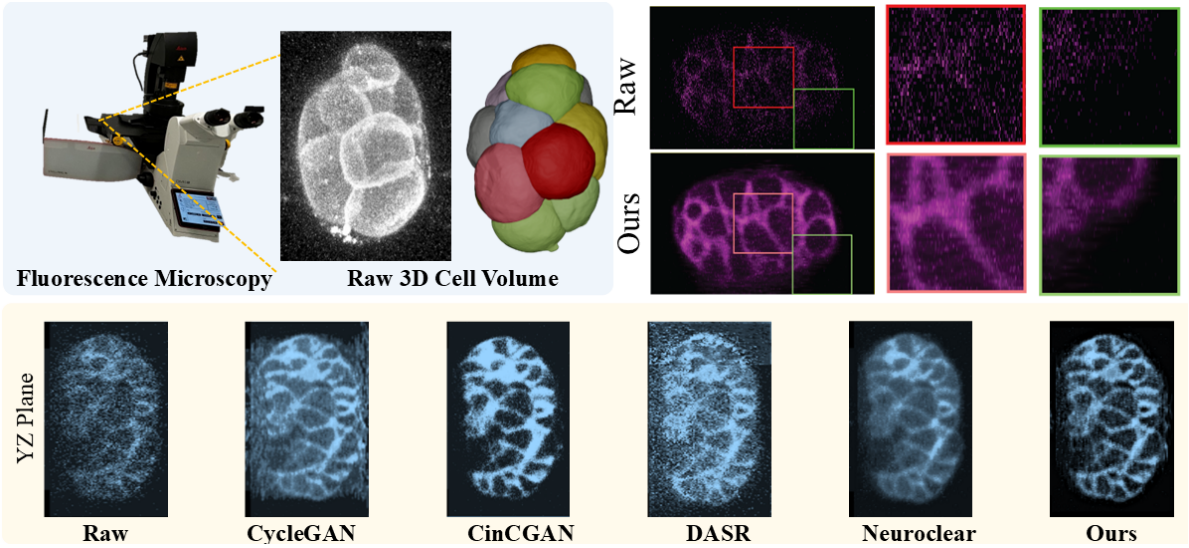


Figure 1. Problem statement and results of our method. **Top-Left**: The fluorescence microscopy can observe long-time live-cell life processing, but face problems like spatially varying noise and anisotropic resolution. **Top-Right**: The raw slice (above) and our results (bottom). **Bottom**: The comparisons among our methods and other SOTA methods.

Abstract

3D fluorescence microscopy is essential for understanding fundamental life processes through long-term live-cell imaging. However, due to inherent issues in imaging principles, it faces significant challenges including spatially varying noise and anisotropic resolution, where the axial resolution lags behind the lateral resolution up to 4.5 times. Meanwhile, laser power is kept low to maintain cell viability, leading to inaccessible low-noise and high-resolution paired ground truth (GT). To tackle these limitations, a dual Cycle-consistent Diffusion is proposed to effectively mine intra-volume imaging priors within 3D cell volumes in an unsupervised manner, i.e., Volume Tells (VTCD), achieving de-noising and super-resolution (SR) simultaneously. Specifically, a spatially iso-distributed denoiser is designed to exploit the noise distribution consistency between adjacent low-noise and high-noise regions within the 3D cell volume, suppressing the spatially varying noise. Then, in light of the structural consistency of the cell volume, a cross-plane global-propagation SR module propagates high-resolution details from the XY plane into adjacent regions in the XZ and YZ planes, progressively enhancing resolution across the entire 3D cell volume. Experimental results on 10 $in\ vivo$ cellular dataset demonstrate high improvements in both de-noising and super-resolution, with axial resolution enhanced from $\sim 430\ nm$ to $\sim 90\ nm$.

[1]Department of Electrical Engineering, City University of Hong Kong, Hong Kong (SAR China)
[2]Centre for Intelligent Multidimensional Data Analysis, Hong Kong Science Park, Hong Kong (SAR China)
[3]Department of Biology, Hong Kong Baptist University, Hong Kong (SAR China)

*Equal contributions
†Correspondence: rogerhzk@gmail.com

1. Introduction

3D fluorescence confocal (FC) microscopy is crucial to reveal detailed volumetric structures within biological samples, which is often inaccessible in traditional 2D imaging.

When used for long-term live-cell observation [9, 18, 19], it provides the 3D context of complex biological processes, thereby making significant contributions to fields such as cell biology, neuroscience, and others by enabling the observation of these processes in their 3D context. However, due to the mechanism of 3D FC imaging, it faces two major limitations, spatially varying noise and anisotropic resolution, as shown in Fig. 1. Meanwhile, to maintain cell viability during long-term observation, the laser power must remain low, making it impossible to obtain high-resolution and low-noise 3D volumes, i.e., paired ground truth (GT) is inaccessible. These limitations severely damage spatial structural information, hindering precise volumetric analysis in biology fields.

As an inherently ill-posed inverse problem, the image SR problem has benefited greatly from the rapid development of deep learning [31, 32, 37]. Most of these methods are supervised, relying on paired HR-LR images, like regression-based deep learning, such as SRCNN [7], FSRCNN [8] and de-noising diffusion probabilistic models (DDPM) [11]. However, in 3D FC microscopy, the structural differences between individual cells during development make it impossible to obtain paired HR-LR images. Common unsupervised deep learning methods, such as learned degradation [1] (like CycleGAN [42]), Cycle-in-Cycle SR [38] (like CinCGAN [38], CycleGAN [42]), and deep image prior [30], fail to effectively exploit the unique imaging priors of 3D FC microscopy, leading to limit SR performance. Furthermore, the noise distribution in 3D FC microscopy varies spatially across different slices, making it challenging for existing deep learning methods to simultaneously perform de-noising and SR for 3D FC microscopy.

Current similar methods primarily focus on 3D FC microscopy under different imaging conditions, addressing either SR for low-resolution images/volumes or de-noising for high-resolution images/volumes [12, 27]. For instance, Huang et al. [13] used images from STEM as ground truth to enhance the resolution of confocal FC microscopy, while studies such as [22, 23, 26] employed self-supervised learning or Circle GAN to improve isotropic resolution restoration for volumetric FC microscopy. These methods either require more resource consumption to obtain high-resolution references or fail to exploit the internal priors of the 3D volume within a self-supervised learning framework.

To address the above limitations, this paper proposes a dual cycle-consistent diffusion (VTCD) for 3D FC microscopy, which leverages two distinct intra-volume imaging priors in 3D cell volumes to achieve simultaneous denoising and SR during training, as described in the title, *Volume Tells*. In the inference stage, the two generators in our trained model serve as the denoiser and the SR module, by inputting the low-resolution, high-noise 3D cell volume into the denoiser and SR module sequentially, a low-noise, high-resolution 3D cell volume can be obtained. Our method models the forward and reverse processes of de-noising and SR, enabling the model to accurately learn the intra-volume imaging priors in 3D cell volumes. To summarize, our key contributions are four-fold:

- Our method exploits intra-volume imaging priors in 3D cell volumes as the forward stage of DDPM, and integrates DDPM into the cycle training, enabling simultaneous denoising and SR for 3D FC microscopy.
- A spatially iso-distributed denoiser model the denoising as a condition diffusion model, which learns the prior distribution of partial low-noise XY plane images and optimizes adjacent relatively high-noise regions, gradually reducing the spatially varying noise throughout the entire 3D cell volume without paired GT in the cycle training.
- A cross-plane global-propagation SR module enables the diffusion model to learn the prior distribution of continuous cell structures in the XY plane, propagating local structural information along the Z-axis within a more global space, continuously refining the super-resolved 3D cell volume without paired GT.
- Live-cell fluorescence datasets under diverse imaging conditions are collected, containing four-dimensional, single shot with great laser and axial imaging with high frequency. Extensive experimental results on these comprehensive datasets demonstrate significant improvements in SNR and resolution with multiple metrics.

2. Related Works

2.1. Deep Learning-based SR

Recent advances in deep learning have propelled the SR task to new heights, with a variety of architectures tailored to improve both image quality and computational efficiency. Convolutional Neural Networks (CNNs) initially laid the foundation for SR with early models like SRCNN and VDSR [7, 15], which demonstrated the effectiveness of end-to-end learning for image SR. Based on these, recent CNN-based SR methods have integrated attention mechanisms to focus on salient image regions, leading to marked improvements in SR performance and visual fidelity [6, 39]. The introduction of GANs has brought transformative advancements to SR, notably enhancing perceptual quality by enabling models to generate visually plausible details [16, 35]. More recently, Transformer-based models have further revolutionized SR, leveraging self-attention mechanisms to capture long-range dependencies and improve representation capabilities [2–5, 20, 41]. IPT [2] demonstrated the potential of Transformers pre-trained on large-scale datasets for SR. Subsequently, SwinIR [20] introduced a hierarchical framework Swin Transformer [21], which efficiently processes images at multiple scales and has achieved state-of-the-art (SOTA) results. Latest SR models include using

graphs [29], diffusion model [17, 36, 40], dictionary learning [33], but these methods either require a large amount of paired HR-LR data or their performance is not accurate and stable enough.

2.2. SR of 3D Fluorescence Microscopy

3D fluorescence microscopy SR has become a crucial research focus, aiming to address the inherent limitations of traditional fluorescence microscopy. DFCAN [24] leveraged differences in frequency content across various features to learn hierarchical representations of high-frequency information related to diverse biological structures. Park et al. [23] utilized a 3D optimal transport-driven CycleGAN network (OT-CycleGAN) for the isotropic recovery of various volumetric imaging data. Self-Net [22] utilized a self-learning strategy to enhance the axial resolution of the anisotropic raw data by employing high-resolution lateral images from the same dataset as target references. Huang et al. [13] designed a two-channel attention network (TCAN) to learn the mapping from low-resolution images to high-resolution ones. Qu et al. [26] proposed a Self-inspired Noise2Noise module (SN2N) that employs self-supervised data generation and a self-constrained learning process to enhance various super-resolution imaging modalities. A method most comparable to this study is [25], which utilized a SimCLR-like approach to simultaneously denoise and perform super-resolution on volumetric fluorescence microscopy. However, their method can't exploit the imaging prior of 3D cell volume, failing to achieve accurate and stable results. The comparison is presented in detail in experiments.

3. Methodology

3.1. Overview of VTCD

The image super-resolution (SR) problem, an inherently ill-posed inverse problem, has been significantly advanced by deep learning, though most methods rely on paired high- and low-resolution images, which are unavailable in 3D fluorescence microscopy due to structural differences between cells. Existing unsupervised methods, such as learned degradation and Cycle-in-Cycle SR, fail to leverage the unique priors of 3D fluorescence microscopy, limiting SR performance.

As shown in Fig. 2, since 3D FC microscopy is essentially an unsupervised problem, inspired by [23], the VTCD is designed based on the cycle-in-cycle framework. For effectively mining intra-volume imaging priors, the VTCD will model the de-noising and SR processes as two conditional diffusion de-noising processes (CDDP).

Given a 3D cell volume $V \in \mathbf{R}^{H\times W\times C}$, the tth XY/XZ/YZ plane slices I^t_{xy}, I^t_{xz} and I^t_{yz}, where H, W and C is X-axis, Y-axis and Z-axis, H and W is high resolution, Z is low resolution. The VTCD consists of two generators G_r and G_f, and two discriminators D_r and D_f, r and f notes the forward stage and reverse stage respectively. The spatially iso-distributed denoiser and cross-plane global-propagation SRM are noted as $SDD(\cdot)$ and $SRM(\cdot)$.

In the forward stage, we utilize the characteristics of 3D FC microscopy to construct a special slicing method, which is equivalent to the forward stage of both CDDPs. For the DDPM forward stage, the original diffusion model gradually adds Gaussian noise to the image I_0 at $t \in \{1, \dots, T\}$ to obtain the noisy image I_T. For de-noising in our case, in fluorescence microscopy, the light source is incident along the Z-axis, causing the XY plane slices along the Z-axis, $\{I^0_{xy}, I^1_{xy}, \dots, I^T_{xy}\}$ with increasing noise. Therefore, we perform continuous slicing along the Z-axis, obtaining a set of XY plane slices, as shown in the top-left part of Fig. 2, which is equivalent to the forward stage, viewed as

$$q(I^t_{xy}|I^{t-1}_{xy}) = N(\sqrt{1-\beta_t}I^{t-1}_{xy}; \beta_t I_d) \tag{1}$$

Similar to the slicing above, for the forward stage of SR, the 3D volume slicing will produce three sets, XY/XZ/YZ plane slices, $\{I^0_{xy}, I^1_{xy}, \dots, I^T_{xy}\}$, $\{I^0_{xz}, I^1_{xz}, \dots, I^T_{xz}\}$, $\{I^0_{yz}, I^1_{yz}, \dots, I^T_{yz}\}$, as shown in the bottom-left part of Fig. 2, can be viewed as

$$q(I^t_{tmp}|I^{t-1}_{xy}) = N(\sqrt{1-\beta_t}I^{t-1}_{xy}; \beta_t I_d) \tag{2}$$

where I^t_{tmp} is the tth slice of XZ/YZ planes.

In the reverse stage of the de-noising CDDP, the spatially iso-distributed denoiser guides the generative trajectory via the semantic content of the corresponding slice. In the reverse stage of the SR CDDP, the cross-plane global-propagation SRM maps the XY-plane slice features to the corresponding coordinates of a generated 3D grid, guiding the SR generative trajectory. Next, the spatially iso-distributed denoiser and the cross-plane global-propagation SR module will be introduced in detail with the description of the reverse stage.

3.2. Spatially Iso-Distributed Denoiser

In the reverse stage of de-noising CDDP, the key is to control the generative trajectory to obtain the LN (lower noise) $\hat{I}^t_{xy}$ rather I^0_{xy}. Thus, the spatially iso-distributed denoiser uses the semantic content of HN (higher noise) slice I^t_{xy} to guide the generative trajectory at the reverse stage.

In the reverse stage, the original diffusion model denoises a latent sample $\hat{I}^T_{xy}$ to a data sample I^0_{xy} as

$$I^{t-1}_{xy} = \frac{1}{\sqrt{1-\beta_t}}(I^t_{xy} - \frac{\beta_t}{\sqrt{1-\alpha_t}}; \epsilon^\theta_t(I^t_{xy})) + \sigma_t z_t \tag{3}$$

However, we aim to obtain the low noise Tth slice I^T_{xy} rather than I^0_{xy}, the original DDPM is unable to address this

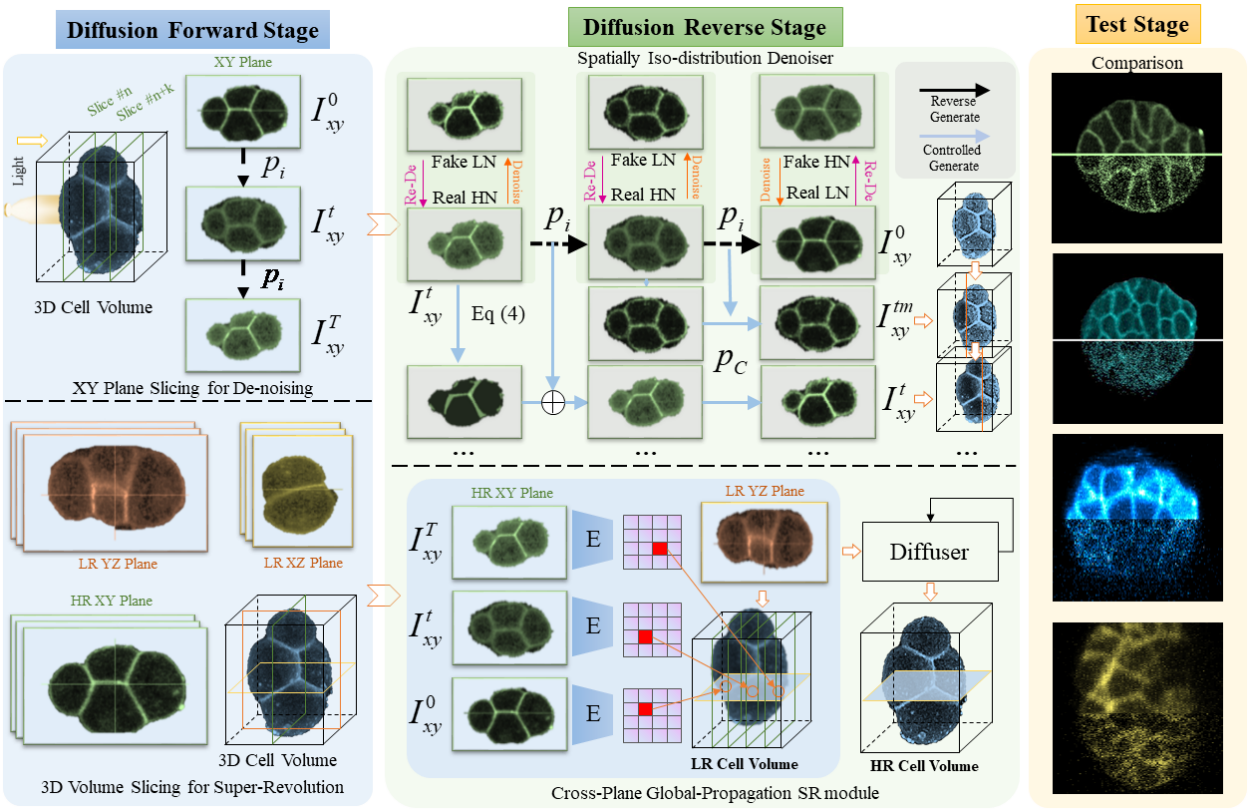


Figure 2. The overall framework of the proposed method. **Left**: two targeted slicing strategies are modeled as the forward stage of diffusion model via the imaging prior for de-noising and SR 3D cell volume. **Middle**: the spatially iso-distributed denoiser model is a conditional diffusion model, which progressively reduces the noise of 3D cell volume along the Z-axis, and the cross-plane SRM enables the conditional diffusion model propagates the content distribution of the HR XY plane to the adjacent 3D volume space, eventually the whole 3D space. **Right**: The comparison of our results (above) and original slices (below). Compared to the original slice, our results show significantly reduced noise, and previously unobservable structural information becomes clear and discernible.

issue. Thus, We guide the generative trajectory in the reverse process of I^t_{xy} using the shape and semantic information of I^t_{xy}, as shown in Fig. 2. Following [28], We define the semantic consistency of the latent encoding with respect to the semantic information of I^t_{xy} as:

$$d(I^t_{xy}, n) = n^T I^t_{xy} \tag{4}$$

where n is the unit normal vector of the hyperplane. Then the latent encodings can be edited by $x'_t = x_t + \lambda d(x_t, s_t)$. At the same time, we use the contextural loss to better control the generated results, which will be described in detail in supplementary materials. The reverse stage can be presented as

$$x'_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(x'_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}; \epsilon^\theta_t(x'_t)) + \sigma_t z_t \tag{5}$$

The distribution prior progressively de-noises XY plane slices along the Z-axis. Through continuously editing this reverse stage at every slice of $\{I^0_{xy}, I^1_{xy}, \ldots, I^T_{xy}\}$, the 3D cell volume can be denoised with the imaging prior, and enhance the semantic information of every slice to preserve its true cell shape. This progressive de-noising effectively reduces the spatially varying noise across the entire 3D cell volume.

3.3. Cross-Plane Global-Propagation SR module

For the reverse stage of SR's CDDP, in our case, the high resolution of the XY plane serves as a guide for the SR of the XZ and YZ planes, ultimately resulting in an HR 3D cell volume. However, most existing diffusion models rely on ample HR 3D data to supervise this process and get the 3D cell volume. In our case, such ample HR 3D cell volumes are evidently unavailable, leaving only the 2D HR slices of the XY plane as a resource.

Thus, in the reverse stage of SR, the cross-plane global-propagation SRM is designed to propagate the HR information from the XY plane to the surrounding 3D cell volume. This module generates intermediate 3D-aware features, then progressively refines and reconstructs the

3D cell volume, ensuring that the anisotropic resolution is restored while maintaining consistency with the high-resolution structural details from the XY plane.

Given a 3D cell volume $V \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the X, Y, and Z axes, respectively, with H and W corresponding to high-resolution dimensions and C corresponding to a low-resolution dimension, we extract three slices of the volume: the XY, XZ, and YZ plane slices, denoted as I^t_{xy}, I^t_{xz}, and I^t_{yz}. Following the approach outlined in [10, 14], we first project the 3D coordinates x^V_{hwc} of each grid element (h, w, c) from the 3D cell volume onto each corresponding XY plane slice I_{xy}. A fixed ResNet32 is then applied to extract the 2D features f^t_{hwc} for each slice, with t indicating the specific XY plane slice.

The result is a generated grid $\bar{V} \in \mathbb{R}^{d\times H\times W\times C}$, where each grid element is associated with a corresponding 2D feature. To enhance the resolution of $\bar{V}$, an Accumulator MLP, A_{acc}, is employed to aggregate the surrounding spatial information and produce a new grid element. Specifically, A_{acc} takes as input the features f_{mno} from neighboring grid elements, where $m = h \pm$ shift, $n = w \pm$ shift, and $o = c \pm$ shift, with shift set to 1 in our implementation. The output of A_{acc} is a feature vector $\theta \in \mathbb{R}^k$, where k corresponds to the total number of possible shifts. This process can be interpreted as

$$f_{hwc} = \sum_{j}^{k} \theta_j f^j_{mno} \tag{6}$$

Thus, an HR grid $\bar{V} \in \mathbb{R}^{d\times H\times W\times C}$ is generated by cross-plane propagating the HR information of XY plane. In the reverse stage of SR, we only need to update I^t_{xz} and I^t_{yz} from XZ/YZ plane by

$$I'^t_{xz} = I^t_{xz} + SL\,(V) \tag{7}$$

where $SL(\cdot)$ represents the operation that forms a slice from the features of I^t_{xz} at the corresponding spatial location $\bar{V}$, and overlays it onto the original LR slice. Subsequently, Eq. 5 is applied in the reverse stage.

The SR process leverages the structural consistency of the 3D cell volume to propagate HR information from the XY plane to adjacent spatial regions along the Z-axis. Through the step-by-step application of the aforementioned process to XZ/YZ plane slices, the high resolution of the XY plane is globally propagated throughout the 3D volume. Thus, the SR process ensures that HR information from the XY plane is effectively extended to the entire 3D cell volume, achieving anisotropic super-resolution with structural consistency.

Due to space limitations, the training strategy and loss are included in the supplementary materials.

4. Experiments

4.1. The Novel Dataset and Experimental Setup

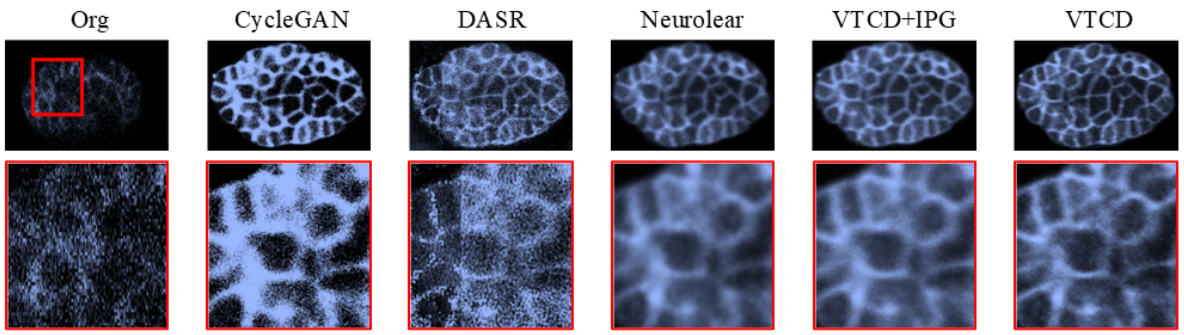
Dataset	Image Num.	Axial Res.	Laser Intens.
UnpairedTrain	22560	94	Middle
FullrefXY-1	1028	94	Low
FullrefXY-2	584	92	Middle
FullrefXY-3	456	30	High
NorefZ-1	887	94	Middle
NorefZ-2	1098	200	High
NorefZ-3	487	30	High

Table 1. The training and benchmark evaluation dataset we provided for all FC image processing (cell analyses especially).

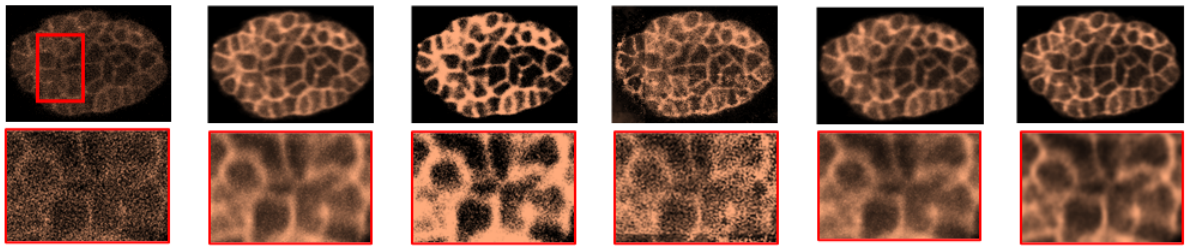
All imaging datasets were collected from live-cell embryos, using either the Leica SP5 II or Stellaris confocal microscopy systems. The images were subsequently processed and analyzed with Leica LAS X software and ImageJ. For 4D data acquisition, to maintain embryo viability, light intensity was kept low by using a scanning speed of 8000 Hz with a water immersion objective, which resulted in significant degradation, noise, and spatially anisotropic signals. High-frequency axial and single-shot images were captured with a higher laser power for observation purposes only, not for data curation. The imaging was divided into time blocks, with each segment corresponding to a 60-time-point interval. The z-axis compensation ranged from 0.5% to 3% for the 488 nm laser and from 20% to 95% for the 594 nm laser.

As demonstrated in Table 1, we provide a total of 600 new 3D fluorescence cell volumes (22560 lateral 2D images, UnpairedTrain) for training and evaluation which was not available before. With higher weights for late cell stage embryos, we randomly generate unpaired 11000 images as a training dataset. To construct a benchmark for cell fluorescence imaging super-resolution application, from these 2D images, we generate 3 evaluation datasets with full reference (FullrefXY-1, FullrefXY-2, FullrefXY-3) and other 3 with no reference (NorefZ-1, NorefZ-2, NorefZ-3) for a comprehensive comparison between our method and existing SOTA unsupervised learning methods. Full reference datasets are constructed on XY plane images, while the input images are downscale 4 times (128×178) and the reference (evaluating GT) is the original XY images (512×712).

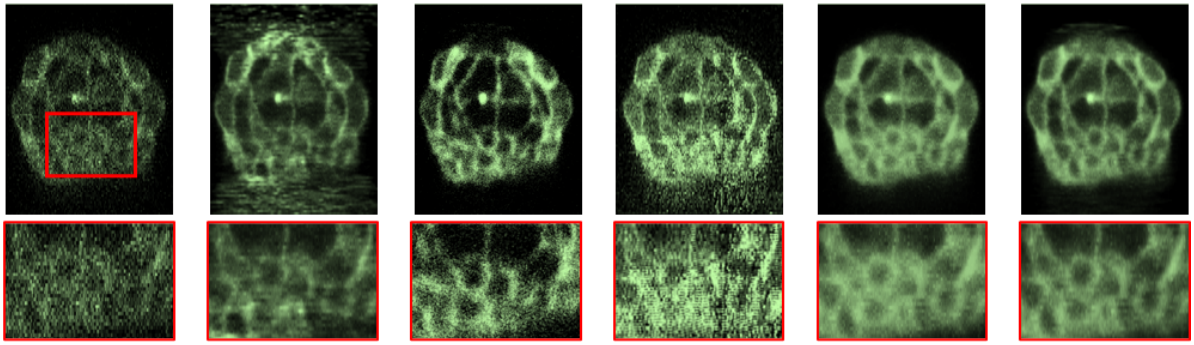
The model training used mCherry-labeled (cell membrane stained) fluorescence images as inputs, without the HR GT images serve as the learning targets. The real 4D volumes are in the shape of 512×712×94 for keeping cells healthy. The training was optimized using the Adam optimizer, starting with an initial learning rate of 5×10^{-3} and a weight decay of 1×10^{-5}, along with the AMSGrad vari-



(a) Comparison of results in the XY plane near the upper Z-axis region, showing differences in detail preservation and clarity.



(b) Comparison of results in the XZ plane, highlighting differences in structure and resolution.



(c) Comparison of results in the YZ plane, illustrating differences in structure and resolution.

Figure 3. The qualitative comparison of the cell fluorescence dataset from three extremely different aspects. These results are from different cell and different life time, thus there are different number and shape of cells.

ant for improved gradient descent. The proposed model and other methods are all trained on 20 CPUs (Intel(R) Xeon(R) Platinum 8457C), and 3 NVIDIA A100 GPUs.

4.2. Qualitative De-noising and Super-resolution Comparisons

To further evaluate the effects of our method, we compared VTCD and VTCD-based methods with other existing unsupervised learning models [3, 23, 29, 34, 38, 42]. On the same cell membrane fluorescence GT (Table 1, unpaired images in low and high resolution respectively). Comprehensive qualitative comparisons demonstrate the significant improvement from VTCD in Fig. 3 and Supplementary Information. The de-noising effects perform on all XY, YZ, and XZ planes while the super-resolution module solves the isotropic low-resolution issues in the YZ and XZ planes. It is obvious that in the raw fluorescence images, the cell membrane signals are almost impossible to recognize in the Z-direction (Fig. 3). These optical and physical restrictions hinder the resolution of the third axis Z. We demonstrate that our method VTCD dramatically interpolates the information at the YZ and XZ plane and makes the Z-direction viewing clear enough for human eye recognition (Supplementary Note).

Besides the great enhancements from VTCD for the raw images, comparing other existing good unsupervised learning methods, our proposed method has outperformed them qualitatively. No matter from local texture restoration, global noise elimination, or Z-axis super-resolution interpolation, VTCD has the best interpretation for the cell structural signals. For example, unlike the chaotic and global noise in CycleGAN and Neuroclear [23, 42] second and

Dataset	Method	Years	Full Reference Metrics				No Reference Metrics		
			PSNR↑	SSIM↑	LPIPS↓	VIF↑	NIQE↓	PIQE↓	NRQM↑
FullrefXY-1	CycleGAN	2018	32.5422	0.6200	0.4576	0.0533	21.2918	47.3730	4.4334
	CinCGAN	2019	31.1730	0.5997	0.3204	0.0537	20.9956	42.2347	4.0328
	DASR	2021	32.2968	0.6197	0.4876	0.0376	20.1543	45.1351	4.5786
	Neuroclear	2022	30.5578	0.5798	0.3122	0.0487	21.8716	48.1651	4.4882
	VTCD+HAT	2023	32.3669	0.6236	**0.3069**	0.0363	**16.8470**	49.9810	4.3840
	VTCD+IPG	2024	**32.5718**	**0.6338**	**0.3040**	**0.0630**	**16.9656**	**41.5012**	**5.0619**
	VTCD	Ours	**32.5719**	**0.6357**	0.3221	**0.0653**	17.0562	**41.0117**	**4.6480**
FullrefXY-2	CycleGAN	2018	38.5692	0.9766	0.7076	0.0372	21.6533	48.2666	4.3256
	CinCGAN	2019	35.3893	0.5896	0.5311	0.0297	20.8319	52.3552	3.5771
	DASR	2021	38.5214	0.9612	0.4682	0.0430	20.4517	47.2133	3.5114
	Neuroclear	2022	36.7488	0.8405	0.5654	0.0281	19.5124	52.9998	3.5871
	VTCD+HAT	2023	**39.8612**	**0.9826**	**0.3648**	0.0369	20.5301	48.6379	4.5130
	VTCD+IPG	2024	37.9154	0.9519	**0.3138**	**0.0700**	**18.9704**	**46.4613**	**4.6895**
	VTCD	Ours	**40.7851**	**0.9866**	0.3708	**0.0566**	**15.6277**	**44.7752**	**4.5877**
FullrefXY-3	CycleGAN	2018	33.6425	0.7860	0.2557	0.0516	20.9541	54.5335	4.0513
	CinCGAN	2019	33.4772	0.5198	0.3766	0.0731	19.4749	53.5507	3.6956
	DASR	2021	34.2431	0.8066	0.2777	0.0234	16.8766	50.2188	3.8611
	Neuroclear	2022	33.5124	0.7339	0.3665	0.0486	16.2343	53.9991	3.8761
	VTCD+HAT	2023	34.1687	0.8053	0.2077	**0.0989**	16.0564	**48.9741**	4.0874
	VTCD+IPG	2024	**34.8173**	**0.8081**	**0.1995**	0.0954	**15.3609**	**48.6838**	**4.4972**
	VTCD	Ours	**34.9251**	**0.8137**	**0.1986**	**0.1106**	**15.9362**	53.4817	**4.2382**

Table 2. The comparison results on 3 full reference datasets. Multiple metrics are applied in the experiments for fair and comprehensive evaluation. ↑ means higher value is better performance; ↓ indicates that smaller value is better.

fourth columns in Fig. S1, the results from VTCD are much cleaner. In DSAR's results, the super-resolution for YZ and XZ eliminates some signals when the cell densities are high (at the left part of the embryos) but VTCD could balance these issues in a good way for detailed observations.

4.3. Evaluation on Dataset with Full and No Reference Metrics

The quantitative evaluation of our VTCD model was conducted against 6 SOTA unsupervised learning methods across both full-reference and no-reference metrics on various datasets, including FullrefXY-1, FullrefXY-2, FullrefXY-3, and NorefZ-1 to NorefZ-3. The results consistently demonstrate the superior performance of our proposed VTCD in FC images' de-noising and super-resolution.

For the FullrefXY-1, FullrefXY-2, and FullrefXY-3 datasets, our VTCD model consistently outperforms competing methods in both PSNR and SSIM metrics. In FullrefXY-1, VTCD achieves a PSNR of 32.5719 (∼6.59% over Neuroclear) and an SSIM of 0.6357 (∼2.6% over CycleGAN), surpassing prior models such as CinCGAN [38] and DASR [34]. The model also performs competitively across perceptual metrics, maintaining an LPIPS score of 0.3221 and achieving a lower PIQE of 41.0117 compared to other baseline models, indicating a significant reduction in perceived image distortions. In FullrefXY-2, VTCD attains a PSNR of 40.7851 (∼5.9% over DASR) and an SSIM of 0.9866 with the best. VTCD also excels in reducing perceptual artifacts, as evidenced by its NIQE score of 15.6277, which is notably lower than that of most competitors, showcasing its robustness in preserving image details and clarity in super-resolution tasks. The model's performance is further validated by its NRQM score of 4.5877, highlighting a well-balanced approach to enhancing perceptual quality. These metrics reflect VTCD's effective handling of both image fidelity and perceptual quality (Fig. 4).

In the axial super-resolution tasks, where ground truth is unavailable, the NorefZ-1, NorefZ-2, and NorefZ-3 datasets were used for evaluation using no-reference metrics. VTCD demonstrates notable strengths, particularly in PIQE and NRQM scores, which are indicators of image quality and perceptual realism. In NorefZ-1, VTCD achieves a PIQE of 34.7048, reflecting a 9.8% improvement over VTCD+IPG [29], and maintains a competitive NIQE score of 15.9237. Similarly, in NorefZ-2, VTCD attains the highest NRQM of 5.7390, emphasizing the model's ability to achieve reduced noise and maintain image quality across challenging

Dataset	Method	No Reference Metrics		
		NIQE↓	PIQE↓	NRQM↑
NorefZ-1	CycleGAN	21.1589	45.6363	3.5341
	CinCGAN	16.4999	49.9127	2.9472
	DASR	16.9833	49.5444	3.5225
	Neuroclear	16.8188	48.0947	3.7322
	VTCD+HAT	16.1298	43.9811	**4.3409**
	VTCD+IPG	**15.7458**	**38.5549**	**5.1336**
	VTCD	**15.9237**	**34.7048**	4.0482
NorefZ-2	CycleGAN	16.2938	45.3978	4.3399
	CinCGAN	16.8824	49.5337	4.8908
	DASR	14.5744	47.6895	5.3026
	Neuroclear	16.3213	48.1098	4.9018
	VTCD+HAT	13.9311	**46.5909**	5.2468
	VTCD+IPG	**12.8717**	46.6903	**5.6680**
	VTCD	**13.4206**	**46.5737**	**5.7390**
NorefZ-3	CycleGAN	18.3970	66.9309	2.4942
	CinCGAN	14.8976	73.5437	2.3943
	DASR	12.8424	76.5410	3.6288
	Neuroclear	19.3217	69.1047	2.2472
	VTCD+HAT	**12.2402**	**60.2010**	**5.7724**
	VTCD+IPG	**12.1711**	62.7601	4.8728
	VTCD	12.4983	**53.2591**	**6.3571**

Table 3. The comparison results on 3 no reference datasets. Three no reference metrics (DNN pre-trained model) are used in calculating the performance of the super-resolution images. ↑ means higher value is better performance; ↓ indicates that smaller value is better.

no-reference scenarios.

Overall, these quantitative evaluation results confidently demonstrate the superior performance of VTCD across both full-reference and no-reference metrics. The model not only excels in enhancing image clarity and reducing perceptual artifacts but also maintains robustness in preserving fine details, demonstrating its effectiveness across diverse datasets.

4.4. Ablation Studies

Module Name	S-D	C-S	DcL	PSNR	SSIM
Base Model	✗	✗	✗	31.9431	0.6211
+SD	✓	✗	✗	32.3147	0.6298
+SD+CS	✓	✓	✗	**32.4987**	**0.6308**
+SD+DcL	✓	✗	✓	32.4237	0.6278
+SD+CS+DcL	✓	✓	✓	**32.5719**	**0.6357**

Table 4. The ablation studies on 2 proposed modules and novel loss functions with FullrefXY-1 dataset. SD denotes SID-Denoiser, CS denotes CPGP-SRM, and DcL denotes Dual Cycle-consistent Losses.

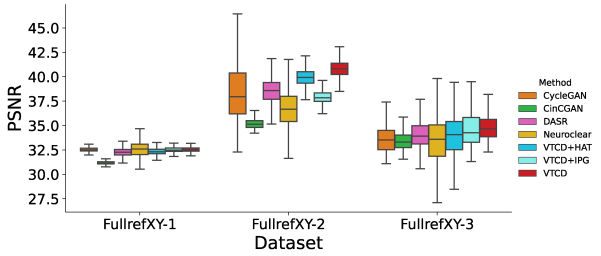


Figure 4. Comparison of PSNR values across different methods in 3 datasets (under multiple imaging conditions), highlighting the statistical performance variations.

Without SID-Denoiser: When the SID-Denoiser was removed, we observed a significant reduction in the denoising capability. The PSNR dropped by approximately 1.16%, and SSIM decreased from 0.6298 to 0.6211, indicating that the model struggled to suppress spatially varying noise effectively. This highlighted the importance of spatially distributed noise handling in achieving high-quality 3D reconstructions.

Without CPGP-SRM: Removing the CPGP-SRM led to a decline in super-resolution quality, especially along the z-axis. The PSNR decreased by 0.57%, and the visual inspection showed a lack of clarity in the XZ and YZ planes. This confirmed the necessity of global-propagation to enhance resolution anisotropically and propagate XY-plane high-resolution details.

Without Dual Cycle-consistent Losses: The exclusion of cycle-consistent loss resulted in both reduced PSNR and SSIM, suggesting less effective feature consistency across different planes. The model's PSNR dropped by 0.46%, and the resulting images showed visible structural inconsistencies, particularly along the boundaries of cellular structures.

5. Conclusion

The proposed VTCD leverages the unique imaging priors of 3D fluorescence microscopy, and within the framework of cycle training, it models de-noising and SR as two conditional diffusion models without paired GT. This approach addresses the limitations of current deep learning-based methods in de-noising and SR for 3D fluorescence microscopy, while simultaneously achieving de-noising and SR of 3D cell volumes. Through the dual objectives of de-noising and SR, VTCD not only restores the suppressed details in the Z-axis but also enhances the structural consistency across 3D cell volumes. This dual capability is particularly beneficial for understanding fundamental life processes in long-term time-lapse live-cell studies. Moreover, the VTCD framework is designed to generalize across various experimental conditions, making it a versatile tool for improving the quality of 3D fluorescence microscopy data.

Acknowledgement

This work is supported by Hong Kong Innovation and Technology Commission (ITC) (InnoHK Project CIMDA), Hong Kong Research Grants Council (RGC) (Project 11204821), and the Institute of Digital Medicine of City University of Hong Kong (Project 9229503). The authors have no competing interests or financial conflicts to disclose.

References

[1] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12310, June 2021. 2

[3] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, June 2023. 2, 6

[4] Zheng Chen, Yulun Zhang, Jinjin Gu, yongbing zhang, Linghe Kong, and Xin Yuan. Cross aggregation transformer for image restoration. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25478–25490. Curran Associates, Inc., 2022. 2

[5] Haram Choi, Jeongmin Lee, and Jihoon Yang. N-gram in swin transformers for efficient lightweight image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2071–2081, June 2023. 2

[6] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11057–11066, 2019. 2

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):295–307, 2016. 2

[8] Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 391–407, Cham, 2016. Springer International Publishing. 2

[9] Guoye Guan, Zelin Li, Yiming Ma, Jianfeng Cao, Ming-Kin Wong, Lu-Yan Chan, Hong Yan, Chao Tang, and Zhongying Zhao. Quantitative cell morphology in c. elegans embryos reveals regulations of cell volume asymmetry. *bioRxiv*, 2023. 2

[10] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4709, 2021. 5

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 2

[12] Yiwei Hou, Wenyi Wang, Yunzhe Fu, Xichuan Ge, Meiqi Li, and Peng Xi. Multi-resolution analysis enables fidelity-ensured deconvolution for fluorescence microscopy. *eLight*, 4(1):14, Aug 2024. 2

[13] Boyi Huang, Jia Li, Bowen Yao, Zhigang Yang, Edmund Y. Lam, Jia Zhang, Wei Yan, and Junle Qu. Enhancing image resolution of confocal fluorescence microscopy with deep learning. *PhotoniX*, 4(1):2, Jan 2023. 2, 3

[14] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18423–18433, 2023. 5

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[16] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[17] Jie Long Lee, Chen Li, and Gim Hee Lee. Disr-nerf: Diffusion-guided view-consistent super-resolution nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20561–20570, June 2024. 3

[18] Zelin Li, Zhaoke Huang, Zhen Zhu, Sicheng You, Zhongying Zhao, and Hong Yan. Directional and topological transformer with topology priors for 4d cellular image segmentation. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 2902–2908, 2024. 2

[19] Zelin Li, Dongying Xie, Yiming Ma, Cunmin Zhao, Sicheng You, Hong Yan, and Zhongying Zhao. Deep learning-based enhancement of fluorescence labeling for accurate cell lineage tracing during embryogenesis. *Bioinformatics*, 40(11):btae626, 10 2024. 2

[20] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1833–1844, October 2021. 2

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin trans-

former: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 2

[22] Kefu Ning, Bolin Lu, Xiaojun Wang, Xiaoyu Zhang, Shuo Nie, Tao Jiang, Anan Li, Guoqing Fan, Xiaofeng Wang, Qingming Luo, Hui Gong, and Jing Yuan. Deep self-learning enables fast, high-fidelity isotropic resolution restoration for volumetric fluorescence microscopy. *Light: Science & Applications*, 12(1):204, Aug 2023. 2, 3

[23] Hyoungjun Park, Myeongsu Na, Bumju Kim, Soohyun Park, Ki Hean Kim, Sunghoe Chang, and Jong Chul Ye. Deep learning enables reference-free isotropic super-resolution for volumetric fluorescence microscopy. *Nature Communications*, 13(1):3297, Jun 2022. 2, 3, 6

[24] Chang Qiao, Di Li, Yuting Guo, Chong Liu, Tao Jiang, Qionghai Dai, and Dong Li. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature Methods*, 18(2):194–202, Feb 2021. 3

[25] Chang Qiao, Yunmin Zeng, Quan Meng, Xingye Chen, Haoyu Chen, Tao Jiang, Rongfei Wei, Jiabao Guo, Wenfeng Fu, Huaide Lu, Di Li, Yuwang Wang, Hui Qiao, Jiamin Wu, Dong Li, and Qionghai Dai. Zero-shot learning enables instant denoising and super-resolution in optical fluorescence microscopy. *Nature Communications*, 15(1):4180, May 2024. 3

[26] Liying Qu, Shiqun Zhao, Yuanyuan Huang, Xianxin Ye, Kunhao Wang, Yuzhen Liu, Xianming Liu, Heng Mao, Guangwei Hu, Wei Chen, Changliang Guo, Jiaye He, Jiubin Tan, Haoyu Li, Liangyi Chen, and Weisong Zhao. Self-inspired learning for denoising live-cell super-resolution microscopy. *Nature Methods*, 21(10):1895–1908, Oct 2024. 2, 3

[27] Zafran Hussain Shah, Marcel Müller, Tung-Cheng Wang, Philip Maurice Scheidig, Axel Schneider, Mark Schüttpelz, Thomas Huser, and Wolfram Schenck. Deep-learning based denoising and reconstruction of super-resolution structured illumination microscopy images. *Photon. Res.*, 9(5):B168–B181, May 2021. 2

[28] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 4

[29] Yuchuan Tian, Hanting Chen, Chao Xu, and Yunhe Wang. Image processing gnn: Breaking rigidity in super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24117, June 2024. 3, 6, 7

[30] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[31] Chenwei Wang, Siyi Luo, Jifang Pei, Yulin Huang, Yin Zhang, and Jianyu Yang. Crucial feature capture and discrimination for limited training data sar atr. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204:291–305, 2023. 2

[32] Chenwei Wang, Renjie Xu, Yulin Huang, Jifang Pei, Chuan Huang, Wenqi Zhu, and Jianyu Yang. Limited-data sar atr causal method via dual-invariance intervention. *IEEE Transactions on Geoscience and Remote Sensing*, 2025. 2

[33] Longguang Wang, Juncheng Li, Yingqian Wang, Qingyong Hu, and Yulan Guo. Learning coupled dictionaries from unpaired data for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25712–25721, June 2024. 3

[34] Longguang Wang, Yingqian Wang, Xiaoyu Dong, Qingyu Xu, Jungang Yang, Wei An, and Yulan Guo. Unsupervised degradation representation learning for blind super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10581–10590, 2021. 6, 7

[35] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1905–1914, October 2021. 2

[36] Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25796–25805, June 2024. 3

[37] Zhihao Wang, Jian Chen, and Steven C. H. Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3365–3387, 2021. 2

[38] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018. 2, 6, 7

[39] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[40] Qingping Zheng, Ling Zheng, Yuanfan Guo, Ying Li, Songcen Xu, Jiankang Deng, and Hang Xu. Self-adaptive reality-guided diffusion for artifact-free super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25806–25816, June 2024. 3

[41] Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, and Qibin Hou. Srformer: Permuted self-attention for single image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12780–12791, October 2023. 2

[42] 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 (ICCV)*, Oct 2017. 2, 6