Learning Large-Factor EM Image Super-Resolution with Generative Priors

Jiateng Shou[1] Zeyu Xiao[1] Shiyu Deng[1] Wei Huang[1] Peiyao Shi[3]
Ruobing Zhang[3,2] Zhiwei Xiong[1,2] Feng Wu[1,2,†]

[1]MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences

shoujt@mail.ustc.edu.cn {zwxiong,fengwu}@ustc.edu.cn

Abstract

As the mainstream technique for capturing images of biological specimens at nanometer resolution, electron microscopy (EM) is extremely time-consuming for scanning wide field-of-view (FOV) specimens. In this paper, we investigate a challenging task of large-factor EM image super-resolution (EMSR), which holds great promise for reducing scanning time, relaxing acquisition conditions, and expanding imaging FOV. By exploiting the repetitive structures and volumetric coherence of EM images, we propose the first generative learning-based framework for large-factor EMSR. Specifically, motivated by the predictability of repetitive structures and textures in EM images, we first learn a discrete codebook in the latent space to represent high-resolution (HR) cell-specific priors and a latent vector indexer to map low-resolution (LR) EM images to their corresponding latent vectors in a generative manner. By incorporating the generative cell-specific priors from HR EM images through a multi-scale prior fusion module, we then deploy multi-image feature alignment and fusion to further exploit the inter-section coherence in the volumetric EM data. Extensive experiments demonstrate that our proposed framework outperforms advanced single-image and video super-resolution methods for $8\times$ *and* $16\times$ *EMSR (i.e., with 64 times and 256 times less data acquired, respectively), achieving superior visual reconstruction quality and downstream segmentation accuracy on benchmark EM datasets. Code is available at* https://github.com/jtshou/GPEMSR.

1. Introduction

Electron microscopy (EM) is a commonly used imaging technique in life sciences to investigate the ultrastructure of cells, tissues, organelles, and macromolecular complexes, which captures images of biological specimens at nanometer resolution. However, high-quality EM image acquisition typically requires a strict and time-consuming process, involving careful adjustments of beam current, aperture size, and detector settings. This process may take up to years to scan wide field-of-view (FOV) specimens. For example, Zheng *et al.* [57] spent approximately 16 months to acquire a ∼106TB whole-brain dataset of an adult drosophila melanogaster. The long acquisition time greatly limits the application of EM imaging in analyzing complete biological structures in large specimens, such as neuron connections in mammalian brains.

Image super-resolution (SR), which is capable of restoring high-resolution (HR) images from their corresponding low-resolution (LR) observations, has the potential to revolutionize EM imaging by allowing for faster and less restrictive data acquisition, while also providing high-quality images with a wide field of view. By applying SR to EM images (shorted as EMSR hereafter), the capturing time can be significantly reduced, and the strict capturing conditions can be relaxed. By deploying a simple ResNet-based UNet model, Fang *et al.* [13] have demonstrated the promising performance of EMSR for $4\times$ magnification (*i.e.*, with 16 times less data acquired). However, achieving even larger-factor EMSR to further reduce capturing time remains challenging. This is in accordance with existing methods [5, 18, 30, 35, 50, 58] for natural images, which can achieve satisfactory results for up to $4\times$ magnification, but fail to meet the demand for larger factors.

On the other hand, recent advances in generative models, such as ChatGPT and diffusion-based models [9, 16, 21, 43], reveal powerful capability in automatic content generation, including natural languages and images. This motivates us to consider the EMSR task from a generative perspective. Especially, compared with natural images that possess diverse structures and textures, EM images often exhibit repetitive structures and textures due to the predictability of imaging specimens, making it more suitable to leverage generative learning for accurate reconstruction. In

[†] Corresponding author.

this paper, we propose a novel deep learning-based framework tailored to the challenging task of large-factor EMSR, by 1) exploiting the repetitive structures and textures in EM images with generative cell-specific priors learned from HR EM images, and 2) exploiting the inter-section coherence in the volumetric EM data by aggregating features learned from multiple consecutive images.

Specifically, our framework explores cell-specific priors using a VQGAN-Indexer network, consisting of VQGAN [12] and our proposed latent vector indexer. We first learn a discrete codebook to represent the distribution of HR EM images in the latent space. The codebook captures both structure and texture information, while the decoder establishes relationships between latent vectors and image patches. We then train a latent vector indexer to acquire the corresponding latent vectors and integrate the indexer with the codebook and the decoder for generating HR EM images. By treating the generation process as an indexing task, we can match LR EM images with their corresponding HR feature representations from the latent space, thereby obtaining priors solely derived from HR EM images.

To maintain reconstruction quality while prioritizing downstream segmentation accuracy, we propose a Multi-Scale Prior Fusion (MPF) module for incorporating the above learned cell-specific priors in EMSR. We use the VQGAN-Indexer output as reference images and learn a mask for fusing reference features based on the patch-level cosine similarity between LR EM images and corresponding reference images. To fully utilize the latent vectors and relationships learned by the decoder, we use multi-scale reference features from different layers of the decoder with varying resolutions. Following the MPF module, our framework includes two key steps for exploiting inter-section coherence in the volumetric EM data: multi-image feature alignment (along the axial direction) and multi-image feature fusion. To this end, we introduce a Pyramid Optical-flow-based Deformable convolution alignment (POD) module and a 3D Spatial-Attention fusion (3DA) module. The former leverages a pre-trained optical-flow network SPyNet [42] and deformable convolutions [6, 60], while the latter leverages the spatial attention mechanism and 3D convolutions. Both improve reconstruction quality and downstream segmentation accuracy for large-factor EMSR.

In summary, this paper offers the following contributions. 1) We present the first generative learning-based framework for the challenging task of large-factor EMSR. 2) We introduce the VQGAN-Indexer network to explore generative cell-specific prior information from HR EM images. 3) We propose the MPF module to effectively utilize the generative priors while preserving image fidelity with LR observations, followed by the POD and 3DA modules for multi-image feature alignment and fusion. 4) Extensive experiments demonstrate the superiority of our framework in terms of both reconstruction quality and downstream segmentation accuracy for 8× and 16× EMSR.

2. Related Work

Electron microscopy image super-resolution. Existing EMSR methods can be categorized into two types: restoring isotropic volumes from anisotropic ones, *i.e.*, SR along the axial dimension [8, 20], and reconstructing HR images from corresponding LR observations in the lateral dimensions [7, 13, 40, 46, 53]. We focus on the latter task in this paper, while our proposed framework may also apply to the former task. As a pioneering work in the field of EMSR, Sreehari *et al.* [46] introduce a Bayesian framework and utilize a library-based non-local means (LB-NLM) algorithm to achieve up to 16× EMSR without requiring a training process. However, this non-learning-based method limits performance and is not specifically designed for large-factor EMSR. Along the deep learning line, Nehme *et al.* [40] train a fully convolutional encoder-decoder network on simulated data to reconstruct super-resolved images. Hann *et al.* [7] train a GAN model using pairs of test specimens captured from the same region of interest. Xie *et al.* [53] leverage the attention mechanism to capture inter-section dependencies and shared features among adjacent images. Compared to previous EMSR methods, our framework not only utilizes adjacent EM images but also explores and integrates generative cell-specific priors to tackle the challenging task of large-factor EMSR.

Video super-resolution. Video super-resolution (VSR) aims to restore HR frames by leveraging adjacent temporal information in multiple LR frames. To align temporal features, optical flow [3, 5, 26, 44, 49, 52, 54] and deformable convolution [47, 50], have been widely adopted. Recently, transformer-based approaches [4, 36] yield remarkable advancements in VSR, owing to the utilization of diverse attention mechanisms. Inspired by these VSR methods, to exploit the inter-section coherence in the volumetric EM data, we utilize optical-flow networks and deformable convolutions for multi-image feature alignment, and spatial attention mechanisms for multi-image feature fusion.

Generative priors in image restoration. Generative image restoration methods [11, 31–33] employ the priors from the pre-trained generative adversarial network (GAN), such as StyleGAN [24] and BigGAN [2], to approximate the natural image manifold and synthesize high-quality images. Given the superior performance of discrete codebook-based generative methods in semantic image synthesis, structure-to-image, and stochastic super-resolution tasks [12, 48], recent methods explore codebook-based facial priors [17, 59] by leveraging VQGAN [12] for training. Different from these methods, we propose a latent vector indexer to exploit the information contained within the input LR images, and the MPF module to fuse generative priors.

3. Method

3.1. Overview

As illustrated in Figure 1, the goal of large-factor EMSR is to obtain the super-resolved $I_{SR}^0 \in \mathbb{R}^{rH\times rW\times 1}$, given a sequence of $2N+1$ consecutive LR EM images, $I_{LR}^z \in \mathbb{R}^{H\times W\times 1}$, which should be close to the ground truth image $I_{GT}^0 \in \mathbb{R}^{rH\times rW\times 1}$, where $z \in \{-N, -N+1, \cdots, 0, \cdots, N-1, N\}$ and r is the large scale factor. In this paper, we set $N=2$ and $r=8,16$.

To achieve this, we propose a generative learning-based framework consisting of three stages. Stage I involves exploring generative cell-specific priors through the VQGAN model, which identifies a discrete latent space of HR EM images and generates the HR EM image I_{HRref} from the latent space. This latent space is represented using vectors from the VQGAN codebook $\boldsymbol{C}$. In Stage II, we train a latent vector indexer and connect it with the VQGAN codebook $\boldsymbol{C}$ and VQGAN decoder $\boldsymbol{Q}$ obtained from Stage I. This connection allows us to generate the reference HR EM image I_{Ref} and its corresponding multi-scale generative features F_{Ref}^l from the LR EM image I_{LR}. Finally, in Stage III, we fuse the multi-scale generative features F_{Ref}^l using the MPF module, align adjacent image features in the axial direction using the POD module, and fuse the adjacent image features using the 3DA module. We finally use sub-pixel convolution to reconstruct the HR output I_{SR}^0.

3.2. Exploring Generative Cell-Specific Priors

EM images are characterized by their repetitive structures and textures, such as cellular membranes and subcellular organelles. These features offer an opportunity to exploit their regularity and predictability, motivating our exploration of the generative cell-specific prior in EM images for large-factor EMSR. As shown in Figure 1 (a), our framework for exploring generative cell-specific prior exploration consists of two main steps. First, we identify a discrete latent space that represents the features of HR EM images. Then, we generate the HR EM images from this latent space, leveraging its compact and representative nature.

Identifying a discrete latent space. We aim to identify a discrete latent space that represents HR EM images. To achieve this, we utilize an encoder $\boldsymbol{E}$ to parameterize the posterior categorical distribution of HR EM images $q_\phi(\xi \mid o)$, where ξ represents the variable for latent vectors, o represents the variable for HR EM images and ϕ represents the encoder's parameters. Specifically, we quantize the output feature Z_e from $\boldsymbol{E}$ by mapping it to its closest latent vector in $\boldsymbol{C}$ and obtain the quantified feature Z_d and one-hot index of each mapped HR patch, denoted as $s^{m,n}$

$$Z_d^{m,n} = \arg\min_{Z_q[j]} \|Z_q[j] - Z_e^{m,n}\|_2^2,$$
$$s^{m,n}[k] = \begin{cases} 1, & k = \arg\min_j \|Z_q[j] - Z_e^{m,n}\|_2^2 \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where $Z_q[j]$ denotes the j-th latent vector stored in $\boldsymbol{C}$. $Z_e^{m,n}$ denotes the encoder output feature element at position (m,n) within Z_e, while $Z_d^{m,n}$ denotes the quantified feature element at position (m,n) within Z_d. $s^{m,n}$ is a v-dimensional vector and $s^{m,n}[k]$ denotes the k-th element in $s^{m,n}$. The codebook $\boldsymbol{C}$ consists of v latent vectors, each with a dimensionality of d. Thus, the posterior categorical distribution $q_\phi(\xi \mid o)$ is defined as

$$q_\phi(\xi = k \mid o) = \begin{cases} 1, & Z_q[k] = \arg\min_{Z_q[j]} \|Z_q[j] - Z_e^{m,n}\|_2^2 \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

Generating the HR EM image. Given the quantified feature Z_d, we can generate HR EM image I_{HRref} through the decoder. We parameterize the prior distribution of HR EM images $p_\theta(o \mid \xi)$ through the decoder $\boldsymbol{Q}$, where θ is the parameters of the decoder.

The encoder is composed of multiple Res-blocks [19] and convolution blocks for downsampling, while the decoder is composed of multiple Res-blocks and deconvolution blocks for upsampling. The compression patch size for downsampling is set to p. Both the encoder and decoder leverage self-attention mechanisms to enhance generalization quality. We optimize the latent vector $Z_q[j]$ along with the encoder $\boldsymbol{E}$ and decoder $\boldsymbol{Q}$.

3.3. Generating LR Reference

With the parameterization of the latent space using $\boldsymbol{C}$ and the prior distribution of HR EM images using $\boldsymbol{Q}$, we can generate HR EM images given real HR EM images as input. However, in the large-factor EMSR task, only highly degraded LR images are available as input. Hence, we need to map the LR images to their corresponding quantified feature Z_d to utilize the generative cell-prior stored in $\boldsymbol{C}$.

To achieve this, one straightforward approach is to interpolate the LR images and feed them into the encoder [17]. This straightforward approach approximates the posterior categorical distribution of LR EM images $q(\xi \mid o_\downarrow)$ by utilizing the parameterized posterior categorical distribution of HR EM images $q_\phi(\xi \mid [o_\downarrow]_\uparrow)$ with the interpolation operation. Here, $o_\downarrow$ represents the LR image variable, and $[\cdot]_\uparrow$ denotes the interpolation operation. However, in scenarios where significant degradation occurs in LR input images, the interpolation operation struggles to restore the rich details and textures in HR EM images. Consequently, this mismatch leads to discrepancies between the real HR EM image distribution $p(o)$ and the interpolated HR EM image distribution $p([o_\downarrow]_\uparrow)$, thereby resulting in disparities between $q(\xi \mid o_\downarrow)$ and $q_\phi(\xi \mid [o_\downarrow]_\uparrow)$. Despite the loss of fine-grained details in LR images, the partial preservation of information enables the utilization of such details as priors in mapping LR images to their corresponding quantified feature Z_d. To fully utilize these priors, we propose a latent

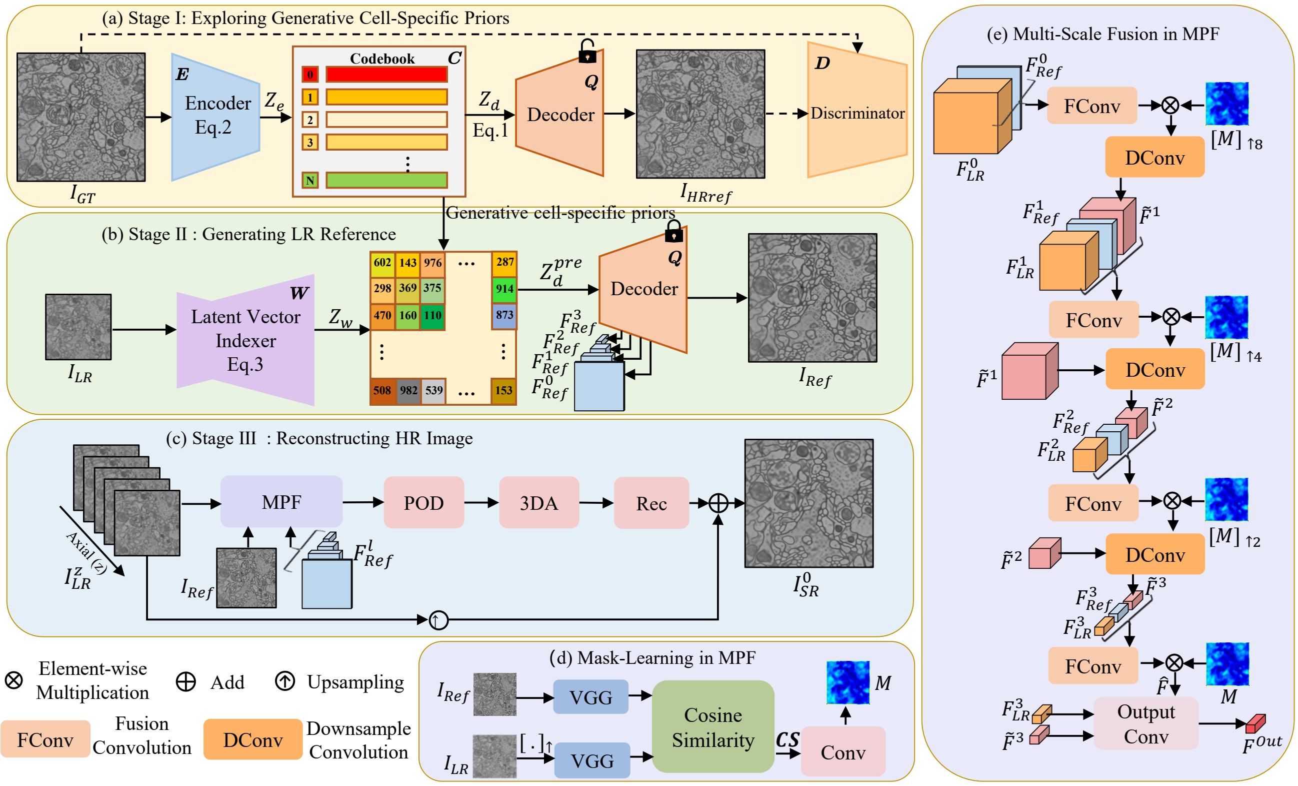

Figure 1. Overview of our framework. The proposed generative learning-based framework consists of three stages. In Stage I, the encoder, the codebook, and the decoder are trained by self-generating the input HR EM image. In Stage II, a latent vector indexer is trained and connected with the codebook and the decoder with fixed parameters to generate the reference HR EM image and multi-scale generative features from the LR EM image. In Stage III, the HR output is reconstructed by fusing multi-scale generative features and exploiting inter-section coherence in volumetric EM data with the POD module and the 3DA module. Rec denotes reconstruction layers composed of convolution layers and pixel shuffle operation.

vector indexer $\boldsymbol{W}$ to predict probabilities of corresponding latent vectors in C given LR EM images as input, as shown in Figure 1 (b). We denote the output of the latent vector indexer as Z_w, where the element at position (m, n) is denoted as $Z_w^{m,n}$ and represents a v-dimensional vector. Then, by selecting the latent vector with the highest probability, we can effectively model the posterior categorical distribution $q(\xi \mid o_\downarrow)$ as

$$q_\varphi\left(\xi = k \mid o_\downarrow\right) = \begin{cases} 1, & k = \arg\max_{i \in \{1,2,3,\cdots,v\}} Z_w^{m,n}[i] \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where φ denotes the parameters of the latent vector indexer $\boldsymbol{W}$. $Z_w^{m,n}[i]$ denotes the i-th element of the vector $Z_w^{m,n}$.

This quantization operation allows us to capture the most representative latent vector that corresponds to the LR image. Then the predicted quantified feature Z_d^{pre} is obtained and fed into the decoder to generate the reference HR EM image I_{Ref} and its corresponding multi-scale generative features F_{Ref}^l through the decoder. Note that all the latent vectors used to generate I_{Ref} and F_{Ref}^l are obtained from the codebook, which contains the generative cell-specific priors of HR EM images.

3.4. Reconstructing HR Image

MPF module. The complete process of reconstructing HR image is depicted in Figure 1 (c). The discrepancies between the generated HR image I_{Ref} and real EM image I_{GT} pose a challenge in achieving accurate multi-scale generative feature fusion. To overcome this challenge, we propose the MPF module, which focuses on identifying and fusing the multi-scale generative features. The MPF module comprises two essential processes: a mask-learning process and a multi-scale fusion process.

The mask-learning mechanism within the MPF module enables us to identify and mask out multi-scale generative features in regions that show significant boundary differences compared to real HR images. As shown in Figure 1 (d), we first embed the interpolated LR image $[I_{LR}]_\uparrow$ and the reference image I_{Ref} from the decoder into the feature space by a pre-trained VGG19 encoder [45]. We extract 16×16 patches from LR feature maps and reference (Ref) feature maps without overlap. Then, We calculate cosine similarity vector $\boldsymbol{CS}$ between LR feature patches and Ref feature patches, where $\boldsymbol{CS_i}$ denotes cosine similarity between the i-th LR feature patch and the i-th Ref feature patch. We refine the mask by employing three 2D convolutions. Finally, we apply a sigmoid function to obtain the mask M with the same spatial resolution as LR features.

To effectively fuse features from different layers of the decoder across L levels, we employ a multi-scale fusion scheme, as illustrated in Figure 1 (e). As the decoder learns the relationship between latent vectors, each layer of the decoder contains varying levels of detailed information about EM images. Additionally, since the decoder is responsible for reconstructing HR EM images, layers closer to the output represent shallower features. Given these considerations, we begin fusing features from F_{Ref}^{0} and its corresponding LR features F_{LR}^{0} with the same resolution and gradually incorporate features with lower resolution

$$\begin{aligned}
\tilde{F}^{1} &= R_D^{0}\left(R_F^{0}\left(\langle F_{LR}^{0}, F_{Ref}^{0}\rangle\right)\otimes[M]_{\uparrow 2^{L-1}}\right),\\
\tilde{F}^{l+1} &= R_D^{l}\left(\langle R_F^{l}\left(\langle \tilde{F}^{l}, F_{LR}^{l}, F_{Ref}^{l}\rangle\right)\otimes[M]_{\uparrow 2^{L-l-1}}, \tilde{F}^{l}\rangle\right),\\
\hat{F} &= R_F^{L-1}\left(\langle \tilde{F}^{L-1}, F_{LR}^{L-1}, F_{Ref}^{L-1}\rangle\right)\otimes M,\\
F^{Out} &= R_O^{L-1}\left(\langle \hat{F}, \tilde{F}^{L-1}, F_{LR}^{L-1}\rangle\right),
\end{aligned} \tag{4}$$

where $\otimes$ denotes element-wise multiplication, $\langle\cdot\rangle$ denotes concatenation operation, and $[\cdot]_{\uparrow}$ denotes interpolation operation. $\tilde{F}^{l}$ denotes output features from downsample convolution at the l-th level. R_D^{l}, R_F^{l}, and R_O^{L-1} denote downsample convolution, fusion convolution, and output convolution, respectively. $\hat{F}$ denotes the output feature of the last fusion convolution. F^{Out} represents the final multi-scale fusion feature.

POD module and 3DA module. To exploit the intersection coherence in volumetric EM data and leverage the correlation between adjacent EM images, we design two modules: the POD module and the 3DA module. The POD module is responsible for aligning the adjacent EM image output features from the MPF module in a coarse-to-fine manner. It utilizes a pyramid architecture and incorporates a pre-trained optical-flow network SPyNet and deformable convolutions to achieve accurate feature alignment. The 3DA module employs spatial attention mechanisms and 3D convolutions to effectively fuse the aligned features. More detailed information about the POD module and 3DA module can be found in the supplementary material.

3.5. Training Strategy

In Stage I, we adopt three loss functions: a reconstruction loss L_{r-I}, a codebook learning loss L_c, and an adversarial loss L_{adv}

$$\begin{aligned}
L_{r-I} &= \left\|I_{\text{HRref}} - I_{\text{GT}}\right\|_1, L_{adv} = D(I_{\text{GT}}) + D(-I_{\text{HRref}}),\\
L_c &= \left\|\text{sg}\left[Z_d\right] - Z_e\right\|_2^2 + \left\|\text{sg}\left[Z_e\right] - Z_d\right\|_2^2,
\end{aligned} \tag{5}$$

where D represents a PatchGAN discriminator [23] and $\text{sg}[\cdot]$ denotes the stop-gradient operation. As Eq. 1 is non-differentiable, we propagate the gradients from the decoder to the encoder [12, 48]. We utilize R1 regularization [39] for GAN training stability. The complete objective of the first stage is

$$L_{s-I} = L_{r-I} + \lambda_c L_c + \lambda_{adv} L_{adv}. \tag{6}$$

In Stage II, we optimize the latent vector indexer $\boldsymbol{W}$ while keeping the parameters of $\boldsymbol{C}$ and $\boldsymbol{Q}$ fixed. The objective of the second stage is

$$L_{s-II} = -\sum_{m,n,i} s^{m,n}[i] \log \frac{\exp\left(Z_w^{m,n}[i]\right)}{\sum_j \exp\left(Z_w^{m,n}[j]\right)}. \tag{7}$$

In Stage III, we train our network with two loss functions: a reconstruction loss L_{r-III} and a multi-Ref fidelity loss L_{Mfid} [28]. The objective of the third stage is

$$\begin{aligned}
L_{r-III} &= \left\|I_{SR}^{0} - I_{GT}^{0}\right\|_1,\\
L_{Mfid} &= \frac{\sum_{z'\in\Omega}\sum_i \delta_i\left(I_{SR}^{0}, I_{Ref}^{z'}\right)\cdot c_{z',i}}{\sum_{z'\in\Omega}\sum_i c_{z',i}},\\
L_{s-III} &= L_{r-III} + \lambda_{Mfid} L_{Mfid},
\end{aligned} \tag{8}$$

where $\Omega = \{-N, -N+1, \cdots, 0, \cdots, N-1, N\}$, $\delta_i(\cdot,\cdot)$ denotes contextual loss in [38] and $c_{z',i}$ is the matching confidence weight based on cosine similarity.

4. Experiments

4.1. Settings

Datasets. We conduct experiments under two different settings to demonstrate the superior performance of our framework in both reconstruction quality and downstream segmentation accuracy. For the first setting, we select the adult drosophila melanogaster brain dataset FAFB [57] and its subset CREMI [14] acquired at $4\times4\times40\text{nm}^3$ resolution. In Stage I and Stage II, we train our framework on a subset (∼38G) of FAFB to obtain a cell-specific prior. In Stage III, we utilize the padded version of the CREMI dataset, excluding the CREMI C subset, to train for large-factor EMSR. We use the cropped version of CREMI C, which has segmentation labels, as the test set consisting of 125 images. We fine-tune two pre-trained segmentation networks, Superhuman [29] and MALA [15], on the first 75 images of CREMI C. Therefore, we test all 125 images for reconstruction evaluation metrics and the last 50 for segmentation evaluation metrics. For the second setting, we select the mouse somatosensory cortex dataset Kasthuri15 [25] and its subset AC3/AC4 dataset [25] acquired at $3\times3\times29\text{nm}^3$ resolution. As the AC3/AC4 datasets are too small, we use the data partitioning strategy from [22] and train three stages on a subset (∼11.2G, *i.e.*, Subset3 in [22]) of Kasthuri15 and AC3. We fine-tune two pre-trained segmentation networks on AC3 and test them on AC4.

Table 1. Quantitative comparison of reconstruction quality and EM image segmentation results on CREMI C for 16× and 8× EMSR. The best and the second-best results are highlighted in **bold** and underline. * Parameters optimized in Stage III.

Methods	Scale	Reconstruction Metrics				Segmentation Metrics				Params/M
		Fidelity		Perceptual		Superhuman		MALA		
		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	VOI↓	ARAND↓	VOI↓	ARAND ↓	
Bicubic	16×	21.5557	0.4801	0.7338	0.4842	6.6627	0.7043	6.5599	0.8830	-
RCAN [56]	16×	23.1238	0.5779	0.5336	0.3429	5.2762	0.6193	2.9353	0.2325	12.9080
SwinIR [35]	16×	23.1098	0.5781	0.5332	0.3416	5.1254	0.4955	2.9302	0.2317	12.1426
BSRN [34]	16×	22.9161	0.5645	0.5453	0.3402	5.5728	0.6809	3.2034	0.2694	0.4719
Real-ESRGAN [51]	16×	22.7631	0.5537	0.5333	0.3134	4.0823	0.4890	3.3806	0.2973	16.6957
EDVR [50]	16×	**23.8894**	**0.6147**	0.5022	0.3182	4.1709	0.4132	2.3229	0.1969	3.4455
BasicVSR [5]	16×	23.3877	0.5924	0.5107	0.3176	4.5825	0.5053	2.5428	0.2115	3.9339
Ours	16×	23.6767	0.6020	**0.4790**	**0.2927**	**2.9639**	**0.3075**	**2.2571**	**0.1786**	5.4627*
Bicubic	8×	25.5897	0.6607	0.4779	0.2982	5.9304	0.8754	2.4748	0.2036	-
RCAN [56]	8×	29.3926	0.7914	0.3729	0.2339	3.3858	0.3288	1.5634	0.1333	12.7603
SwinIR [35]	8×	29.4334	0.7921	0.3724	0.2340	3.3494	0.3267	1.5400	0.1306	11.9949
BSRN [34]	8×	29.1464	0.7832	0.3791	0.2313	3.5928	0.3461	1.5621	0.1315	0.3611
Real-ESRGAN [51]	8×	29.1457	0.7843	0.3774	0.2317	3.5740	0.3363	1.5555	0.1309	16.6957
EDVR [50]	8×	**29.7326**	**0.8016**	0.3642	0.2316	3.1739	0.3159	1.5107	0.1324	3.2978
BasicVSR [5]	8×	29.0828	0.7831	0.3774	0.2305	3.4619	0.3336	1.5879	0.1349	4.0445
Ours	8×	29.7027	0.8002	**0.3568**	**0.2291**	**2.9376**	**0.3076**	**1.5011**	**0.1298**	4.4257*

Table 2. Quantitative comparison of reconstruction quality and EM image segmentation results on AC4 for 16× and 8× EMSR. The best and the second-best results are highlighted in **bold** and underline. * Parameters optimized in Stage III.

Methods	Scale	Reconstruction Metrics				Segmentation Metrics				Params/M
		Fidelity		Perceptual		Superhuman		MALA		
		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	VOI↓	ARAND↓	VOI↓	ARAND ↓	
Bicubic	16×	19.0078	0.2518	0.7758	0.4639	6.9300	0.9688	6.8144	0.9664	-
RCAN [56]	16×	19.5938	0.2981	0.6661	0.3970	6.4535	0.9351	6.4716	0.9294	12.9080
SwinIR [35]	16×	19.5949	0.2956	0.6671	0.3893	6.8796	0.9496	6.5417	0.9322	12.1426
BSRN [34]	16×	19.5613	0.2928	0.6725	0.3914	7.1961	0.9615	6.6168	0.9395	0.4719
Real-ESRGAN [51]	16×	19.4254	0.2893	0.6427	0.3706	6.3853	0.8745	6.0659	0.9140	16.6957
EDVR [50]	16×	**19.9374**	**0.3283**	0.6232	0.3731	4.8178	0.7158	5.0029	0.8282	3.4455
BasicVSR [5]	16×	19.4293	0.2888	0.6747	0.3747	6.8608	0.9598	6.7548	0.9513	3.9339
Ours	16×	19.7468	0.3114	**0.5748**	**0.3584**	**4.3655**	**0.6126**	**4.8362**	**0.8233**	5.4627*
Bicubic	8×	21.0979	0.3720	0.5896	0.3361	6.7694	0.9675	6.1346	0.9157	-
RCAN [56]	8×	22.3167	0.4773	0.5200	0.3162	2.2101	0.2826	2.3571	0.3886	12.7603
SwinIR [35]	8×	22.4909	0.4931	0.5104	0.3213	1.7867	0.2107	2.3635	0.5242	11.9949
BSRN [34]	8×	22.2736	0.4736	0.5231	0.3132	2.2630	0.2860	2.5120	0.3746	0.3611
Real-ESRGAN [51]	8×	22.3439	0.4889	0.5000	0.3127	1.8839	0.2584	2.0680	0.3445	16.6957
EDVR [50]	8×	**22.9314**	**0.5270**	0.4890	0.3176	1.2946	0.2532	1.6486	0.4027	3.2978
BasicVSR [5]	8×	22.2444	0.4757	0.5180	0.3114	2.1716	0.3242	2.5117	0.4953	4.0445
Ours	8×	22.9251	**0.5270**	**0.4654**	**0.3075**	**1.1501**	**0.1377**	**1.5862**	**0.3134**	4.4257*

Metrics. To evaluate the difference between SR results and ground truth, we use various reconstruction metrics. These encompass two fidelity metrics, PSNR and SSIM, alongside two perceptual metrics, LPIPS [55] and DISTS [10]. For EM image segmentation evaluation, we use VOI [41] and ARAND [1] as our metrics. In our segmentation evaluation, we consider both split error and merge error to ensure a comprehensive and accurate assessment.

Training setting. We train the networks in three stages using Adam [27] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$ and employ Cosine Annealing scheme [37]. In our experiments, we set all initial learning rates to 4×10^{-4} except for Stage III on the mouse somatosensory cortex training dataset, where we adjust the initial learning rate to 1×10^{-4} to ensure the convergence of deformable convolution layer. To generate paired data, we use bicubic downsampling to reduce the image resolution by factors of 8 and 16. The LR patch size is set to 16×16 for 16× EMSR and 32×32 for 8× EMSR during Stage III training phase. The compression patch size p is set to 16 based on our ablation studies. We select the weights for our loss functions as follows: $\lambda_c = 10$, $\lambda_{adv} = 0.05$, and $\lambda_{MFid} = 0.01$ for 16× EMSR, and $\lambda_{MFid} = 0.001$ for 8× EMSR. These weights are chosen through empirical experimentation aimed at striking the right balance for optimal performance. We also set the number of latent vectors to $v = 1024$ and their dimension to $d = 512$ in the codebook based on our ablation studies.

4.2. Comparisons with Existing Methods

We compare our framework with (1) single-image SR methods including RCAN [56], SwinIR [35], BSRN [34], and Real-ESRGAN [51], (2) video SR methods including EDVR [50] and BasicVSR [5]. All these models are retrained on the same EM datasets for fair comparisons. The quantitative results of image reconstruction and EM image segmentation are summarized in Table 1 and Table 2.

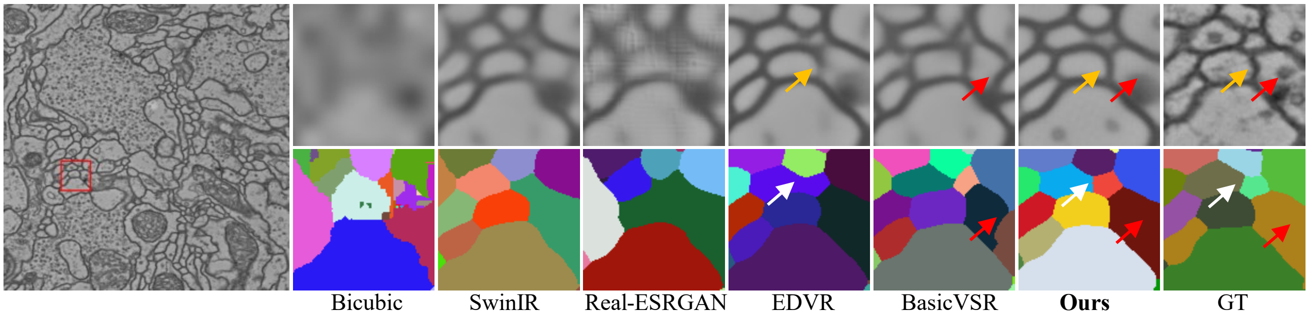

Figure 2. Top: Qualitative comparison for 16× EMSR. Our framework exhibits sharper boundaries and preserves a more complete structure. Bottom: Qualitative comparison for 16× EMSR in terms of segmentation.

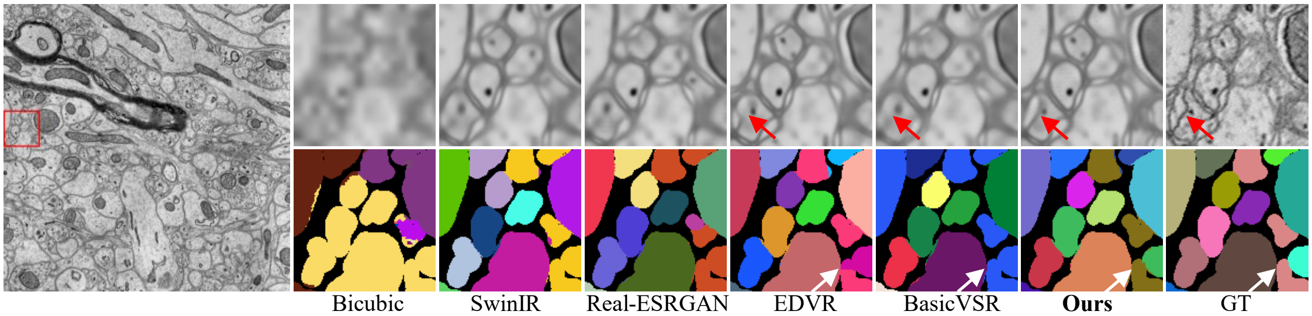

Figure 3. Top: Qualitative comparison for 8× EMSR. Our framework exhibits sharper boundaries and preserves a more complete structure. Bottom: Qualitative comparison for 8× EMSR in terms of segmentation.

Reconstruction quality. While our framework performs the second-best in terms of PSNR and SSIM, slightly behind EDVR, these fidelity metrics do not necessarily align with human perception of image quality. On the other hand, our framework surpasses all other methods in terms of perceptual metrics LPIPS [55] and DISTS [10]. As shown in Figure 2 and Figure 3, our framework produces images with sharper boundaries, richer texture details, and more complete structures for 16× and 8× EMSR. This demonstrates that our cell-specific generative priors and the MPF module can effectively incorporate fine-grained details from HR EM images instead of producing smoothed results at the pixel level. It is noteworthy that although Real-ESRGAN [51] partially alleviates over-smoothing by incorporating the adversarial loss, it introduces noticeable noise-like artifacts in the super-resolved images. Additionally, a general comparison between single-image and video SR methods supports the effectiveness of incorporating adjacent EM images in the context of large-factor EMSR.

EM image segmentation accuracy. We evaluate the accuracy of two pre-trained segmentation networks by using the super-resolved EM images as inputs. Our framework outperforms all other methods with two different segmentation networks, demonstrating exceptional image fidelity in terms of downstream tasks achieved by our framework. Moreover, our framework effectively addresses the issue of over-segmentation caused by excessive smoothing of existing methods, which is particularly prominent in the Superhuman segmentation network (see supplementary material for detail). Visualization examples of segmentation, as shown in Figure 2 and Figure 3, confirm the superiority of our framework over existing methods.

Table 3. Ablation studies on prior exploration and prior indexing trade-off. p denotes the compression patch size.

p	Stage I		Stage II			
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	ΔPSNR	ΔLPIPS
8	35.0731	0.2618	20.1212	0.5240	-14.9519	0.2622
16	28.6876	0.3715	21.7423	0.4996	-6.9453	0.1281
32	23.1879	0.4693	21.5291	0.5024	-1.6588	0.0331
64	20.2832	0.6226	20.1609	0.6263	-0.1223	0.0037

4.3. Ablation Studies

Effectiveness of prior exploration. Table 3 presents a quantitative analysis of the prior exploration process. As the mapping from the HR space to the latent space involves a compression process, the size of the compression patch is a crucial hyperparameter. In Stage I, a smaller compression patch size captures more detailed information and leads to more realistic generated images. We observe a significant improvement in both PSNR and LPIPS as the compression patch size decreases, indicating improvements in reconstruction quality and fidelity.

Trade-off between prior exploration and prior indexing. There exists a trade-off between the quality of generated im-

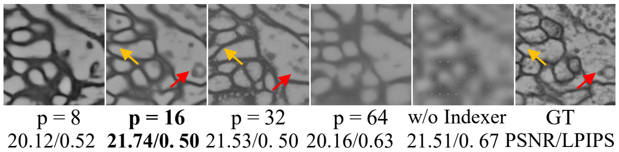

Figure 4. Qualitative comparison of prior indexing results and the effectiveness of the latent vector indexer. p denotes the compression patch size.

Table 4. Ablation studies on the quantity v and dimensions d of latent vectors in the codebook on prior exploration.

v	d	PSNR↑	LPIPS↓
512	512	28.3165	0.3806
1024	256	28.5227	0.3728
1024	512	28.6876	0.3715
2048	512	28.7159	0.3712
1024	1024	28.7574	0.3743

ages in Stage I and the accuracy of predicting latent vectors through the latent vector indexer in Stage II. This trade-off arises from the fact that as the compression patch size decreases and less information is available when feeding LR images into the latent vector indexer in Stage II, the learned posterior distribution of LR EM images $q_\varphi(\xi \mid o_\downarrow)$ deviates further from the true posterior distribution $q(\xi \mid o_\downarrow)$. Consequently, reducing the compression patch size allows the first stage network to preserve more image details, but it poses challenges for the indexing task in Stage II. This can be observed from the changes in ΔPSNR and ΔLPIPS values presented in Table 3, which indicate a decrease in the indexing accuracy as the compression patch size decreases. In consideration of the trade-off between prior exploration and prior indexing, we set $p = 16$, as depicted in Table 3 and Figure 4. Moreover, as illustrated in Figure 4, removing the latent vector indexer and directly utilizing bicubic interpolation along with the decoder in Stage I leads to notably blurry I_{Ref} outcomes.

Impact of the quantity and dimensions of latent vectors in the codebook. We systematically investigate the impact of the quantity and dimensions of latent vectors in the codebook by selecting several sets of representative parameters. The results are summarized in Table 4. From the experimental results, we observe that augmenting v to 2048 or d to 1024 does not yield significant improvements in prior exploration, despite doubling the codebook's size. Conversely, reducing v to 512 or d to 256 results in a noticeable performance decline. Based on these insightful results, we set $v = 1024$ and $d = 512$ for our codebook. This decision strikes an optimal balance between computational efficiency and performance.

Effectiveness of POD module and 3DA module. We evaluate the effectiveness of the proposed POD and 3DA modules for multi-image feature alignment and multi-image feature fusion, respectively. To ensure a fair comparison, we replace these modules with residual blocks with comparable parameters. Table 5 shows the results of our ablation study. We observe that even without incorporating generative prior, the introduction of the POD module or 3DA module leads to improvements in both reconstruction evaluation metrics and segmentation metrics, as demonstrated in row 1 and row 2. Furthermore, when the generative prior is introduced, both the POD module and the 3DA module provide additional gains in reconstruction and segmentation metrics, as shown in rows 3 to 5. Notably, the POD module exhibits a larger improvement compared to the 3DA module, indicating its significance in multi-image feature alignment for large-factor EMSR. These results demonstrate the effectiveness of the proposed POD and 3DA modules in enhancing visual reconstruction quality and segmentation accuracy.

Table 5. Ablation studies of our proposed modules. VOI-S and VOI-M denote segmentation using Superhuman and MALA, respectively.

MPF	POD	3DA	PSNR↑	LPIPS↓	VOI-S↓	VOI-M↓
✗	✗	✗	23.5287	0.5168	4.5912	2.7265
✗	✓	✓	23.7354	0.5063	4.2429	2.3410
✓	✗	✓	23.5422	0.4858	3.1500	2.5144
✓	✓	✗	23.5548	0.4815	3.0193	2.3794
✓	✓	✓	23.6767	0.4790	2.9639	2.2571

Effectiveness of MPF module. The effectiveness of our MPF module is further demonstrated in row 2 and row 5 of Table 5, where it exhibits substantial improvements in perceptual metric LPIPS and segmentation metric VOI. It is important to note that the introduction of L_{Mfid} in Stage III leads to a decrease in PSNR. This observation highlights the inconsistency between PSNR and visual quality as well as segmentation outcomes and reinforces the importance of using perceptual metrics and segmentation metrics to evaluate reconstruction quality.

5. Conclusion

In this paper, we present a generative learning-based framework for large-factor EMSR. By exploring and indexing cell-specific priors using a VQGAN-Indexer network, our framework leverages generative learning to capture the repetitive structures and textures of EM images. The MPF module effectively fuses generative priors obtained from the VQGAN-Indexer network while ensuring the image fidelity. Extensive experiments demonstrate the superiority of our framework in terms of both reconstruction quality and segmentation accuracy for 8× and 16× EMSR.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62021001.

References

[1] Ignacio Arganda-Carreras, Srinivas C Turaga, Daniel R Berger, Dan Cireşan, Alessandro Giusti, Luca M Gambardella, Jürgen Schmidhuber, Dmitry Laptev, Sarvesh Dwivedi, Joachim M Buhmann, et al. Crowdsourcing the creation of image segmentation algorithms for connectomics. *Frontiers in neuroanatomy*, 9:142, 2015. 6

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[3] Jose Caballero, Christian Ledig, Andrew Aitken, Alejandro Acosta, Johannes Totz, Zehan Wang, and Wenzhe Shi. Real-time video super-resolution with spatio-temporal networks and motion compensation. In *CVPR*, 2017. 2

[4] Jiezhang Cao, Yawei Li, Kai Zhang, Jingyun Liang, and Luc Van Gool. Video super-resolution transformer. *arXiv preprint arXiv:2106.06847*, 2021. 2

[5] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Basicvsr: The search for essential components in video super-resolution and beyond. In *CVPR*, 2021. 1, 2, 6

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 2

[7] Kevin de Haan, Zachary S Ballard, Yair Rivenson, Yichen Wu, and Aydogan Ozcan. Resolution enhancement in scanning electron microscopy using deep learning. *Scientific reports*, 9(1):1–7, 2019. 2

[8] Shiyu Deng, Xueyang Fu, Zhiwei Xiong, Chang Chen, Dong Liu, Xuejin Chen, Qing Ling, and Feng Wu. Isotropic reconstruction of 3d em images with unsupervised degradation learning. In *MICCAI*, 2020. 2

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1

[10] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020. 6, 7

[11] Berk Dogan, Shuhang Gu, and Radu Timofte. Exemplar guided face image super-resolution without facial landmarks. In *CVPRW*, 2019. 2

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2, 5

[13] Linjing Fang, Fred Monroe, Sammy Weiser Novak, Lyndsey Kirk, Cara R Schiavon, Seungyoon B Yu, Tong Zhang, Melissa Wu, Kyle Kastner, Alaa Abdel Latif, et al. Deep learning-based point-scanning super-resolution imaging. *Nature methods*, 18(4):406–416, 2021. 1, 2

[14] J Funke, S Saalfeld, DD Bock, SC Turaga, and E Perlman. Miccai challenge on circuit reconstruction from electron microscopy images, 2016. 5

[15] Jan Funke, Fabian Tschopp, William Grisaitis, Arlo Sheridan, Chandan Singh, Stephan Saalfeld, and Srinivas C Turaga. Large scale image segmentation with structured loss based deep learning for connectome reconstruction. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1669–1680, 2018. 5

[16] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 1

[17] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *ECCV*. Springer, 2022. 2, 3

[18] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Recurrent back-projection network for video super-resolution. In *CVPR*, 2019. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[20] Larissa Heinrich, John A Bogovic, and Stephan Saalfeld. Deep learning for isotropic super-resolution from non-isotropic 3d electron microscopy. In *MICCAI*. Springer, 2017. 2

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23(47):1–33, 2022. 1

[22] Wei Huang, Chang Chen, Zhiwei Xiong, Yueyi Zhang, Xuejin Chen, Xiaoyan Sun, and Feng Wu. Semi-supervised neuron segmentation via reinforced consistency learning. *IEEE Transactions on Medical Imaging*, 41(11):3016–3028, 2022. 5

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 5

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[25] Narayanan Kasthuri, Kenneth Jeffrey Hayworth, Daniel Raimund Berger, Richard Lee Schalek, José Angel Conchello, Seymour Knowles-Barley, Dongil Lee, Amelio Vázquez-Reina, Verena Kaynig, Thouis Raymond Jones, et al. Saturated reconstruction of a volume of neocortex. *Cell*, 162(3):648–661, 2015. 5

[26] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *ECCV*, 2018. 2

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[28] Junyong Lee, Myeonghee Lee, Sunghyun Cho, and Seungyong Lee. Reference-based video super-resolution using multi-camera video triplets. In *CVPR*, 2022. 5

[29] Kisuk Lee, Jonathan Zung, Peter Li, Viren Jain, and H Sebastian Seung. Superhuman accuracy on the snemi3d connectomics challenge. *arXiv preprint arXiv:1706.00120*, 2017. 5

[30] Wenbo Li, Xin Tao, Taian Guo, Lu Qi, Jiangbo Lu, and Jiaya Jia. Mucan: Multi-correspondence aggregation network for video super-resolution. In *ECCV*, 2020. 1

[31] Xiaoming Li, Ming Liu, Yuting Ye, Wangmeng Zuo, Liang Lin, and Ruigang Yang. Learning warped guidance for blind face restoration. In *ECCV*, 2018. 2

[32] Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In *ECCV*, 2020.

[33] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *CVPR*, pages 2706–2715, 2020. 2

[34] Zheyuan Li, Yingqi Liu, Xiangyu Chen, Haoming Cai, Jinjin Gu, Yu Qiao, and Chao Dong. Blueprint separable residual network for efficient image super-resolution. In *CVPR*, 2022. 6

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPR*, 2021. 1, 6

[36] Chengxu Liu, Huan Yang, Jianlong Fu, and Xueming Qian. Learning trajectory-aware transformer for video super-resolution. In *CVPR*, 2022. 2

[37] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[38] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018. 5

[39] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *ICML*. PMLR, 2018. 5

[40] Elias Nehme, Lucien E Weiss, Tomer Michaeli, and Yoav Shechtman. Deep-storm: super-resolution single-molecule microscopy by deep learning. *Optica*, 5(4):458–464, 2018. 2

[41] Juan Nunez-Iglesias, Ryan Kennedy, Toufiq Parag, Jianbo Shi, and Dmitri B Chklovskii. Machine learning of hierarchical clustering to segment 2d and 3d images. *PloS one*, 8 (8):e71715, 2013. 6

[42] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 2

[43] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[44] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. Frame-recurrent video super-resolution. In *CVPR*, 2018. 2

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[46] Suhas Sreehari, SV Venkatakrishnan, Katherine L Bouman, Jeffrey P Simmons, Lawrence F Drummy, and Charles A Bouman. Multi-resolution data fusion for super-resolution electron microscopy. In *CVPRW*, 2017. 2

[47] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *CVPR*, 2020. 2

[48] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2, 5

[49] Longguang Wang, Yulan Guo, Li Liu, Zaiping Lin, Xinpu Deng, and Wei An. Deep video super-resolution using hr optical flow estimation. *IEEE Transactions on Image Processing*, 29:4323–4336, 2020. 2

[50] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 1, 2, 6

[51] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *CVPR*, 2021. 6, 7

[52] Zeyu Xiao, Xueyang Fu, Jie Huang, Zhen Cheng, and Zhiwei Xiong. Space-time distillation for video super-resolution. In *CVPR*, 2021. 2

[53] Yaochen Xie, Yu Ding, and Shuiwang Ji. Augmented equivariant attention networks for microscopy image transformation. *IEEE Transactions on Medical Imaging*, 41(11):3194–3206, 2022. 2

[54] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. Video enhancement with task-oriented flow. *International Journal of Computer Vision*, 127:1106–1125, 2019. 2

[55] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6, 7

[56] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 6

[57] Zhihao Zheng, J Scott Lauritzen, Eric Perlman, Camenzind G Robinson, Matthew Nichols, Daniel Milkie, Omar Torrens, John Price, Corey B Fisher, Nadiya Sharifi, et al. A complete electron microscopy volume of the brain of adult drosophila melanogaster. *Cell*, 174(3):730–743, 2018. 1, 5

[58] Shangchen Zhou, Jiawei Zhang, Wangmeng Zuo, and Chen Change Loy. Cross-scale internal graph neural network for image super-resolution. In *NeurIPS*, 2020. 1

[59] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 2

[60] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 2