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

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

Overview

This supplementary document is organized as follows:
Section 1 provides the detailed structure of the encoder, the decoder and the latent vector indexer.
Section 2 provides the detailed structure of the POD module and the 3DA module.
Section 3 provides details of segmentation results.
Section 4 provides more ablation studies.
Section 5 provides the number of parameters of each proposed module.
Section 6 provides visualization of each module of ablation studies.
Section 7 provides discussion about hallucinations.
Section 8 provides more visual comparisons on benchmark datasets.

1. The Structure of the Encoder, the Decoder and the Latent Vector Indexer

We provide the detailed structure of the encoder, the decoder and the latent vector indexer in Figure 1.

Self-attention block. The self-attention blocks in the encoder and decoder employ an attention mechanism to enhance the generation performance. The queries (Q), keys (K), and values (V) are all derived from the same input features. Each of these components is obtained through a 1×1 convolution after applying Group Normalization [10].

We then apply the attention operation

$$\mathrm{Atten}\left(\mathrm{Q},\mathrm{K},\mathrm{V}\right)=\mathrm{V}\cdot\mathrm{softmax}\left(\frac{\mathrm{K}^{T}\mathrm{Q}}{\sqrt{d_k}}\right), \tag{1}$$

where d_k denotes the dimension of the features. Subsequently, we apply a 2D convolution and add the input of the self-attention block. This process yields the output of the self-attention block. The attention mechanism is employed to improve the generation quality in Stage I which is further discussed in Section 4.1.

2. The Structure of the POD Module and the 3DA Module

We provide the detailed structure of the POD module and the 3DA module in Figure 2 and Figure 3, respectively.

POD module. The input of the POD module is adjacent low-resolution EM images I_{LR}^{z} and the output of the MPF module $F^{Out,z}$, where $z \in \{-N, -N+1, \cdots, 0, \cdots, N-1, N\}$. To align multi-image features, we adopt both optical flow and deformable convolution [3, 12]. Taking into account the relationship between deformable convolution offsets and optical flow [2], we incorporate optical flow into the computation of deformable convolution offsets. Following this approach, we utilize SpyNet [7] to compute the optical flow FL_0^1 and FL_0^2. The parameters in the SPyNet network are fixed. Similar to the MPF module, we employ a multi-scale fusion scheme

$$\begin{aligned} O_l^z &= \langle [I_{LR}^z]_{\downarrow 2^l}, [I_{LR}^0]_{\downarrow 2^l}, F_l^{Out,z}, F_l^{Out,0}, FL_l^1, FL_l^2 \rangle, \\ \Delta P_l^z &= Conv_l^2(\langle Conv_l^1(O_l^z), [\Delta P_{l+1}^z]_{\uparrow 2} \rangle), \\ \hat{F}_l^z &= DFconv_l(F_l^{Out,z}, \Delta P_l^z), \end{aligned} \tag{2}$$

where $Conv_l^i$ and $DFconv_l$ denote convolution network and deformable convolution, respectively. $[\cdot]_{\uparrow s}$ denotes upsample interpolaion by a factor of s. $\langle \cdot \rangle$ denotes concatenation operation. $F_0^{Out,z} = F^{Out,z}$. When $l = 2$, $\Delta P_l^z = Conv_l^2(Conv_l^1(O_l^z))$ in Eq. 2. Then we fuse the multi-scale aligned feature with 2D convolutions and refine the alignment results using another deformable convolution with $F^{Out,0}$ as a reference. The output of the POD module is the aligned adjacent EM image features $F^{Align,z}$.

†Corresponding author.

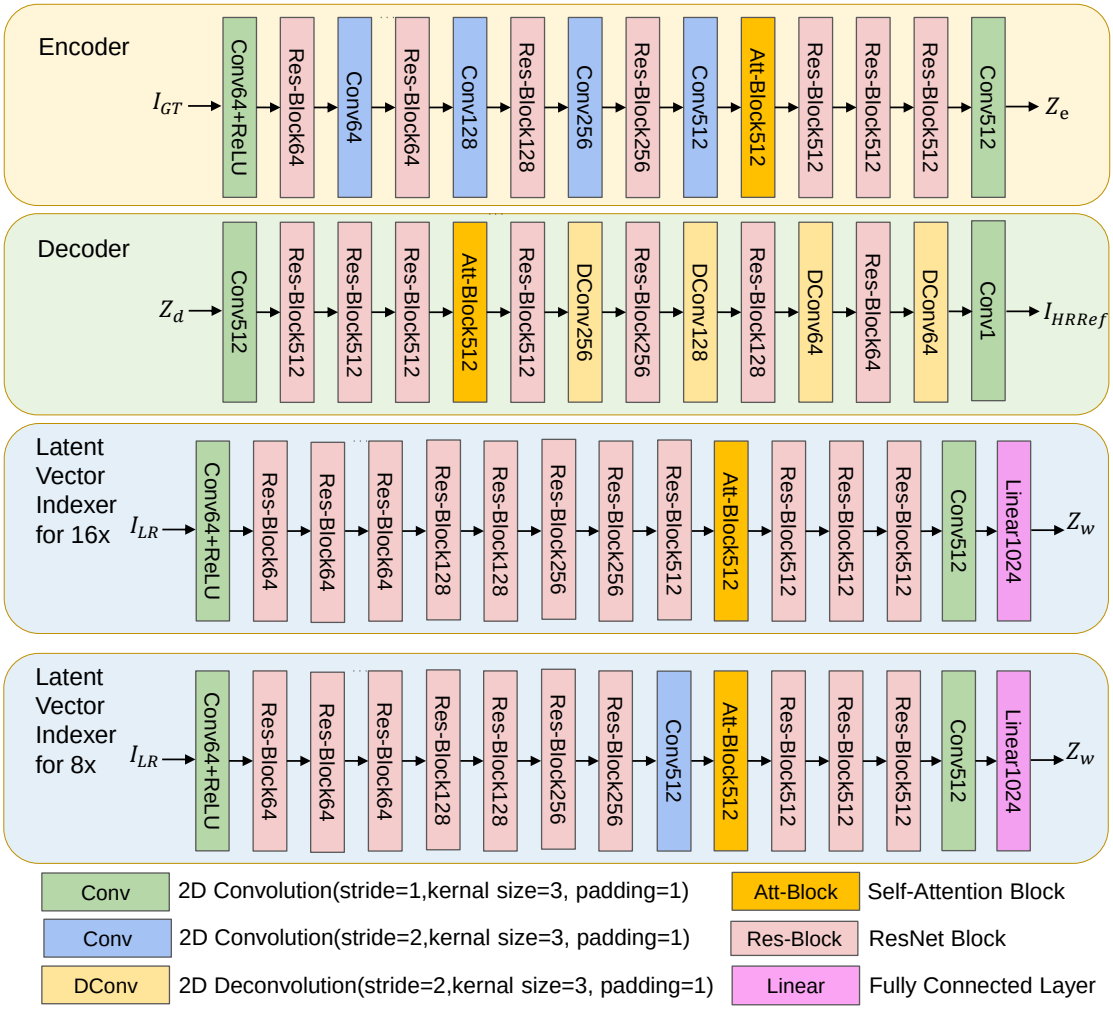

Figure 1. The detailed structure of the encoder, the decoder and the latent vector indexer. The number following the name of each block indicates the number of output channels.

3DA module. The input to the 3DA module consists of the aligned adjacent EM image features $F^{Align,z}$ obtained from the POD module. To eliminate irrelevant features in adjacent images, we first generate masks along the axial direction. Then, leveraging spatial attention mechanisms and 3D convolutions within the 3DA module, we effectively fuse the multi-image features. The resulting output of the 3DA module is the fused adjacent EM image feature denoted as F^{Fused}.

3. Segmentation Details

In Table 1, we report VOI_{split} [6] as an indicator of over-segmentation between our framework and baseline methods including (1) single-image SR methods: RCAN [11], SwinIR [5], BSRN [4], and Real-ESRGAN [9], (2) video SR methods: EDVR [8] and BasicVSR [1]. Our framework achieves the best in terms of VOI_{split}, indicating that our framework effectively addresses the issue of over-segmentation compared to existing methods.

Table 1. Quantitative comparison of VOI_{split} in EM image segmentation on CREMI C for 16× and 8× EMSR. The best is highlighted in **bold**.

Methods	8×		16×	
	Superhuman	MALA	Superhuman	MALA
Bicubic	5.6101	2.0792	5.8078	3.0269
RCAN [11]	3.1635	1.2464	4.6630	2.2929
SwinIR [5]	3.1216	1.2376	4.5242	2.2801
BSRN [4]	3.3587	1.2491	4.8932	2.5038
Real-ESRGAN [9]	3.3346	1.2442	3.1484	2.6901
EDVR [8]	2.9574	1.2235	3.7313	1.8214
BasicVSR [1]	3.2299	1.2508	4.0564	1.9635
Ours	**2.7177**	**1.1993**	**2.4952**	**1.7396**

4. More Ablation Studies

4.1. Effective of Attention Mechanism in the Encoder and the Decoder

To assess the impact of employing the attention mechanism within the encoder and decoder on the quality in Stage I,

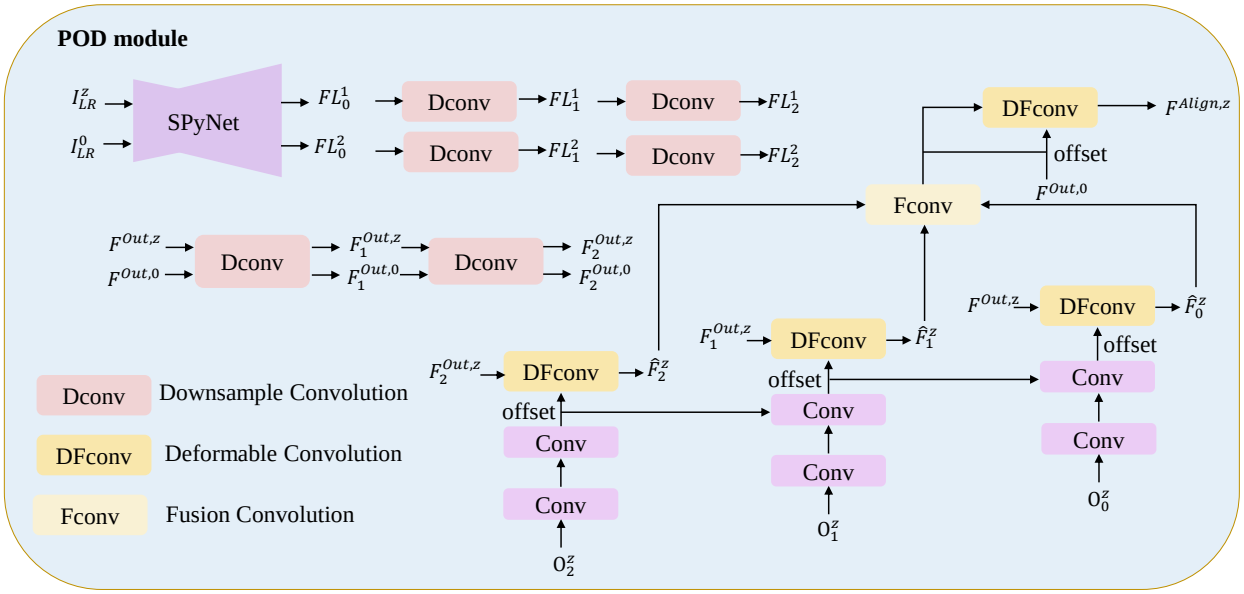

Figure 2. The detailed structure of the POD module. The input of the POD module is adjacent low-resolution EM images I_{LR}^{z} and the output of the MPF module $F^{Out,z}$. The output of the POD module is the aligned adjacent EM image features $F^{Align,z}$. The parameters in the SPyNet network are fixed.

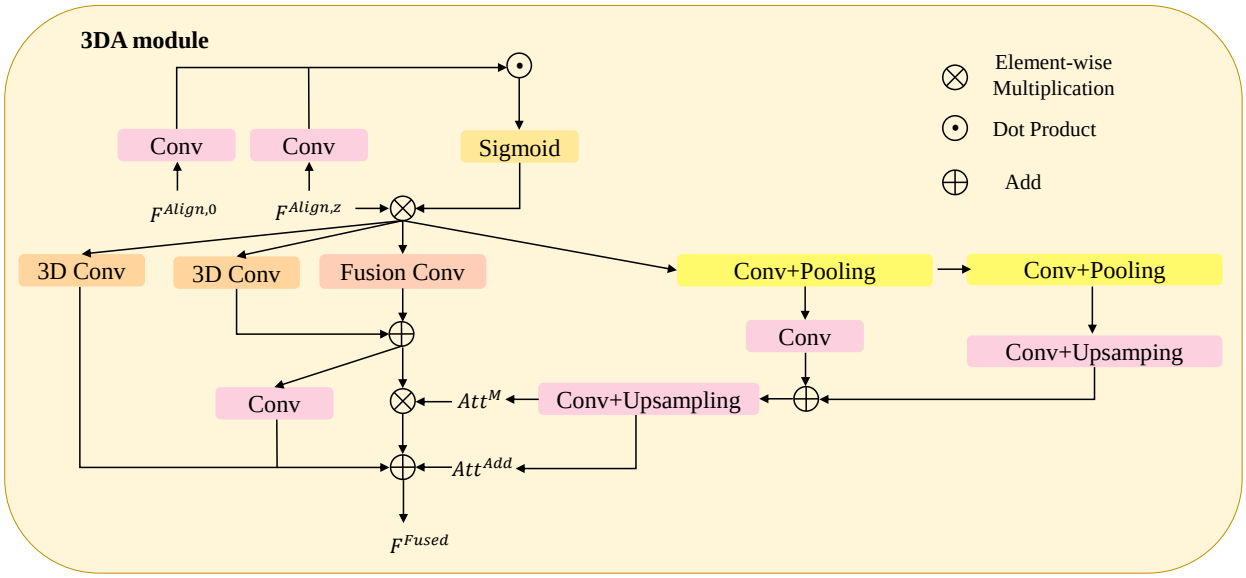

Figure 3. The detailed structure of the 3DA module. The input of the 3DA module is the aligned adjacent EM image features $F^{Align,z}$ from the POD module. The output of the 3DA module is the fused adjacent EM image feature F^{Fused}.

we conduct ablation studies by individually removing the attention mechanism from each component. As shown in Table 2, the inclusion of attention mechanism in both the encoder and decoder yields the best generation quality. This highlights the necessity of employing the attention mechanism in both the encoder and decoder.

4.2. Effectiveness of the Mask in the MPF Module

To assess the indispensability of the mask within the MPF module, we conduct ablation studies by eliminating the mask-learning process. As shown in Table 5, the presence of the mask within the MPF module enhances both reconstruction quality and segmentation accuracy. This highlights the necessity of incorporating the mask in the MPF module.

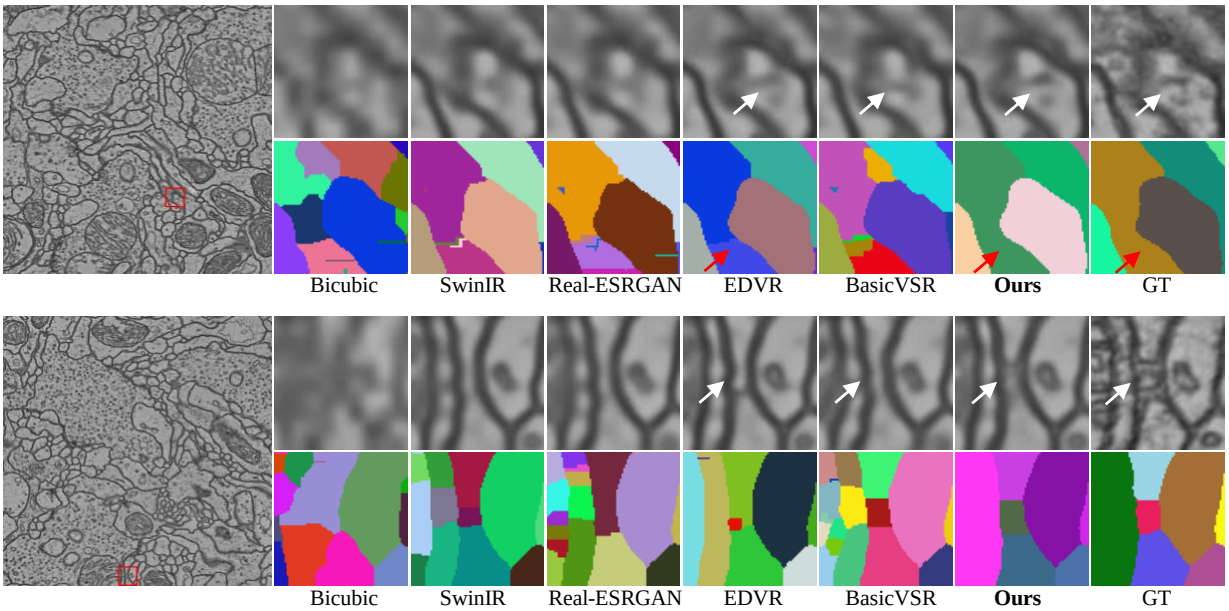

Figure 4. Qualitative comparison for $8\times$ EMSR in terms of reconstruction and segmentation.

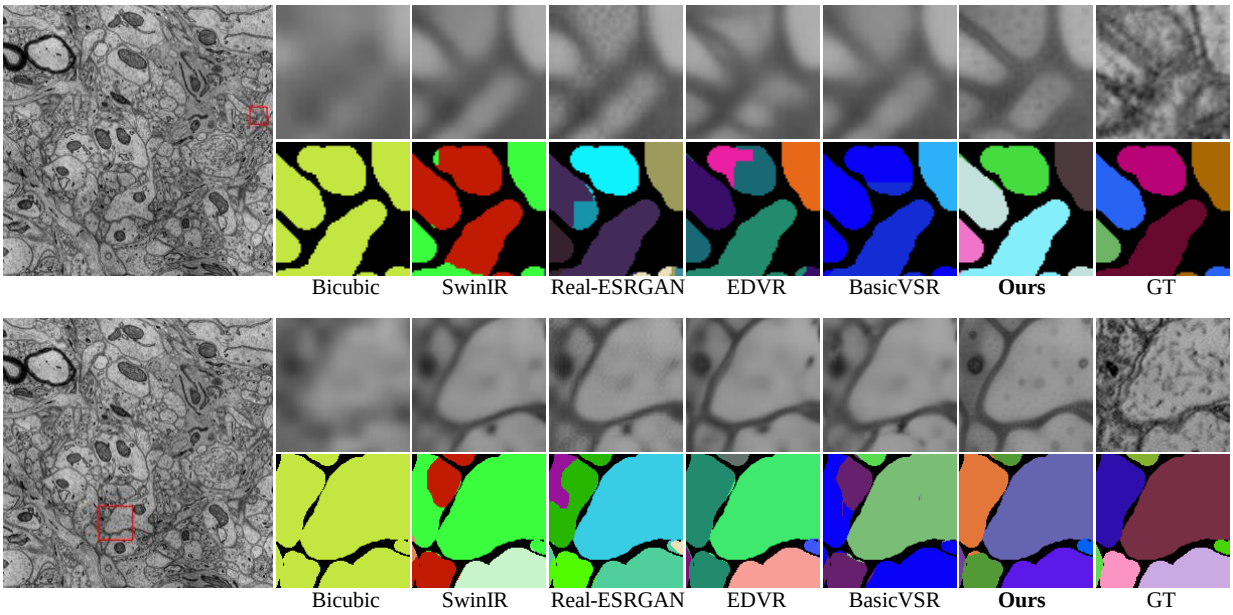

Figure 5. Qualitative comparison for $16\times$ EMSR in terms of reconstruction and segmentation.

Table 2. Ablation studies of the effective of attention mechanism in the encoder and the decoder in Stage I. The best is highlighted in **bold**.

Encoder	Decoder	PSNR↑	LPIPS↓
w/o Att.	w Att.	28.6573	0.3750
w Att.	w/o Att.	28.6483	0.3740
w Att.	w Att.	**28.6876**	**0.3715**

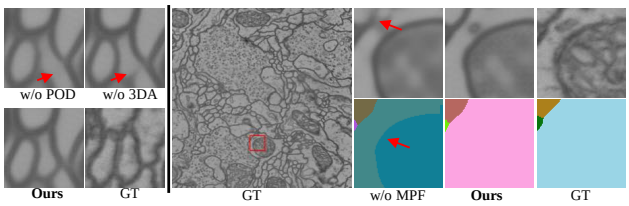

Figure 6. Visualization of each module of the ablation study.

Table 3. Number of parameters of each proposed module for 8× EMSR.

Module	VQGAN-Indexer	MPF	MPF w/o VGG	POD	POD w/o SPyNet	3DA
Params/M	43.3503	21.9235	1.8991	2.8932	1.4529	0.3697

Table 4. Number of parameters of each proposed module for 16× EMSR.

Module	VQGAN-Indexer	MPF	MPF w/o VGG	POD	POD w/o SPyNet	3DA
Params/M	45.8417	21.9276	1.9032	2.8932	1.4529	0.3697

Table 5. Ablation studies of the mask in the MPF Module. The best is highlighted in **bold**.

Method	PSNR↑	LPIPS↓	VOI-Superhuman	VOI-MALA
w/o mask	23.5906	0.4840	3.0886	2.3606
w mask	**23.6767**	**0.4790**	**2.9639**	**2.2571**

5. Number of Parameters

We present the details of the number of parameters of each proposed module in Table 3 and Table 4. It's essential to note that the VQGAN-Indexer network, the VGG network in the MPF module, and the SPyNet network in the POD are fixed in Stage III. Therefore, these parameters are not considered in Table 1 and Table 2 of the paper. Total parameters of all three stages are 72.77M for 16× EMSR and 69.24M for 8× EMSR.

6. Visualization of each Module

We provide visualization of each module of ablation studies. As shown in Figure 6, the absence of the POD/3DA module leads to ineffective utilization of information from adjacent images, resulting in loss of cellular structures. In the absence of the MPF module, the cell boundaries become blurry, making it difficult for the segmentation network to accurately delineate neurons. Consequently, the mitochondrion inside the cell is misclassified as an instance of neuron.

7. Discussion about Hallucinations

Large-scale serial electron microscopy imaging plays a crucial role in reconstructing neural connectomic maps and elucidating the branched morphology of neurons. These datasets contain an enormous amount of structural details, presenting both opportunities and challenges for analysis. While the proposed generative framework may produce unrealistic subcellular structures or compartments, it by and large will not detriment most neural connections (e.g., synapses) or critical morphological features (e.g., stem branchings) as they are relatively large and easily resolvable.

Consequently, the proposed framework is believed to be capable of sustaining the statistical authenticity and validity of the entire dataset. Moreover, the presence of hallucinated details in specific frames can be mitigated by the segmentation network, which examines adjacent images. This additional layer of analysis further diminishes the risk of misleading interpretations, as the segmentation network can effectively identify and exclude hallucinated details.

8. More Visual Results

In Figure 4 and Figure 5, we present more visual comparisons between our framework and baseline methods including (1) single-image SR methods: RCAN [11], SwinIR [5], BSRN [4], and Real-ESRGAN [9], (2) video SR methods: EDVR [8] and BasicVSR [1].

References

[1] 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. 2, 5

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *AAAI*, 2021. 1

[3] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, 2017. 1

[4] 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. 2, 5

[5] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *CVPR*, 2021. 2, 5

[6] 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. 2

[7] Anurag Ranjan and Michael J Black. Optical flow estimation using a spatial pyramid network. In *CVPR*, 2017. 1

[8] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *CVPRW*, 2019. 2, 5

[9] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *CVPR*, 2021. 2, 5

[10] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 1

[11] 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. 2, 5

[12] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *CVPR*, 2019. 1