MFGAN: OCT Image Super-resolution and Enhancement with Blind Degradation and Multi-frame Fusion

Zongqi He*, Zhe Xiao*, Zhuoning Xu*, Yunze Li*, Zelin Song*, Calvin Leighton†, Li Wang*, Shanru Liu*, Shiun Yee Wong†, Wenfeng Huang†, Wenjing Jia†, and Kin-Man Lam*

*The Hong Kong Polytechnic University, Hong Kong, China
†University of Technology Sydney, Sydney, Australia

ABSTRACT

Optical coherence tomography (OCT) is crucial in medical imaging, especially for retinal diagnostics. However, its effectiveness is often limited by imaging devices, resulting in high noise levels, low resolution, and reduced sampling rates, which hinder OCT image diagnosis. This paper proposes a generative adversarial network (GAN) based OCT image super-resolution framework that leverages a blind degradation and Multi-frame Fusion mechanism, namely MFGAN, for retinal OCT image super-resolution. Our method jointly performs denoising, blind super-resolution, and multi-frame fusion, which can reconstruct high quality OCT images without requiring paired ground-truth data. We employ a blind degradation model to handle OCT image degradation and a denoising prior to effectively process noisy inputs. Experimental results on the PKU37 dataset and the VIP Cup 2024 dataset demonstrate that MFGAN excels in both visual quality and quantitative performance, outperforming existing OCT image super-resolution methods.

Keywords: OCT Image Super-resolution, Blind Degradation, Multi-frame Fusion, Image Enhancement

1. INTRODUCTION

Optical coherence tomography (OCT), a non-invasive cross-sectional imaging technique, is essential in various medical applications, particularly in retinal diagnostics.[1] However, the effectiveness of OCT-based diagnosis is hindered by some challenges. OCT images are inevitably corrupted by speckle noise due to the low coherence interferometry imaging modality.[2] Additionally, to expedite the image acquisition process and mitigate the effects of unconscious eye movements during a B-scan, a low sampling rate is often employed. While this approach reduces motion artifacts, it results in low-resolution OCT images.[3] These two issues, *i.e.*, speckle noise and low resolution, significantly degrade the quality of the generated OCT images, which are critical for accurate disease diagnosis. Therefore, it is essential to develop an effective network that achieves image denoising, super-resolution, and enhancement.

Over the past decades, various methods have been proposed to enhance the quality of OCT images. Hardware methods improve the light source and structure of the imaging system, while software post-processing algorithms like BM3D,[4] nonlocal-means,[5] and noise adaptive wavelet thresholding[6] have been developed. Although both approaches can reduce certain types of noise, they often leave residual artifacts or sacrifice structural details. Inspired by the remarkable success of deep learning in various vision tasks,[7–10] deep learning-based methods have demonstrated superior performance in speckle noise reduction, OCT image super-resolution, and detail enhancement, leveraging optimized network architectures.[11]

Various OCT super-resolution techniques have been developed to address low sampling rates and improve image resolution. Traditional interpolation methods often struggle to recover high-resolution (HR) details from noisy and low-resolution (LR) images. To overcome this challenge, Fang et al,[12] introduced sparsity-based simultaneous denoising and interpolation (SBSDI). However, this method is computationally demanding and requires precise alignment of LR and HR images, which is challenging due to the motion artifacts commonly found in OCT imaging. In recent years, deep learning methods

Further author information: (Send correspondence to Kin-Man Lam)
Zongqi He: E-mail: plume.he@connect.polyu.hk
Zhe Xiao: E-mail: xiao-zhe.xiao@connect.polyu.hk
Calvin Leighton: E-mail: calvin.d.leighton@student.uts.edu.au
Wenjing Jia: Email: wenjing.jia@uts.edu.au
Kin-Man Lam: Email: kin.man.lam@polyu.edu.hk

have shown promising advancements in super-resolution, but many still rely on simple bicubic downsampling prior,[13,14] which is not suitable for depicting OCT image degradations. In light of the impressive performance achieved by deep learning models in various vision tasks,[15–17] blind super-resolution has emerged to handle complex and unknown degradations, making it more adaptable to real world scenarios. Some methods utilize explicit degradation models[18] that incorporate factors like blur, downsampling, noise, and JPEG compression to approximate real-world degradations (see Section 2.2.1). However, due to the high variability and complexity of real-world degradation, these methods often struggle in practical applications. Alternatively, others leverage implicit models, such as generative adversarial networks (GANs), to learn degradation patterns directly from data distributions,[19,20] though these approaches may not generalize well to unfamiliar degradation types.

In this work, we propose a novel OCT image processing pipeline that addresses the limitations of existing methods, particularly in handling the complexities of real-world retina OCT images characterized by varying degradation states and significant noise. Our approach incorporates a denoising prior applied to the input image, which helps mitigate the effects of noise before super-resolution. By utilizing a blind degradation module, we avoid the requirement for accurately paired HR and LR images, which are often challenging to obtain. Additionally, we implement a multi-orientation processing strategy where the input image is rotated by multiples of 90 degrees to generate four distinct perspectives. Each version is then processed independently, and the outputs are subsequently fused to produce the final enhanced image. This method effectively enhances OCT images with significant noise and low resolution, refining their detail for clinical diagnosis.

The contributions made in this paper can be summarized as follows:

- We propose a novel framework for OCT retinal image super-resolution that integrates denoising prior, super-resolution, and multi-frame fusion. This approach addresses the unique challenges of noise and limited resolution in OCT images, significantly enhancing their quality.
- We propose a blind super-resolution method. By incorporating a degradation module, the method adapts effectively to the diverse degradation conditions commonly encountered in medical imaging, improving the overall model's robustness for real-world clinical applications.
- Extensive experiments on two benchmark datasets, PKU37[21] and VIP Cup 2024, demonstrate the superiority of the proposed method. The results show that our method achieves leading performance in both visual quality and quantitative metrics, confirming its effectiveness for super-resolution medical image tasks.

2. METHODOLOGY

2.1 Network Architecture

The network of our MFGAN is shown in *Figure 1*. We employ a denoising module used in MPLN to process the LR images, which are then rotated by multiples of 90 ° and passed through four parallel branches, each of which contains three residual-in-residual blocks (denoted as "RR" in *Figure 1*) to produce SR images. These images are then rotated back to their original orientations and fused through multi-frame fusion to generate the final SR output.

2.2 MFGAN for OCT Image Super-Resolution

We propose a GAN-based network for OCT image super-resolution, which is trained with a blind degradation model.[22] The high-resolution images are degraded by the blind degradation model to produce estimated LR images.

2.2.1 Blind Degradation Model

Traditional degradation models can be characterized by a blur kernel, downsampling, and noise, which can be expressed as $y = (x \otimes k) \downarrow_s + n$, where the LR image y is obtained by convolving the HR image x with a blur kernel k, followed by downsampling with a scale factor s, and finally adding additive white Gaussian noise (AWGN) n. Different from the traditional degradation model, we model the key factors, *i.e.*, blur kernel, downsampling, and noise in a more sophisticated way. Specifically, the blur kernel is modeled with two Gaussian blur operations, following the approach proposed in a prior work;[22] the downsampling includes bilinear and bicubic interpolations; the noise is modeled by AWGN, JPEG compression noise, and processed camera sensor noise. Specifically, we perform two consecutive blur operations, each using an isotropic Gaussian kernel or an anisotropic Gaussian kernel, termed as B_i and B_{ani} with a certain probability. For interpolation, we randomly choose between bilinear and bicubic interpolations. Considering OCT images are obtained from imaging devices, we model the noise with processed camera sensor noise $N_c s$.

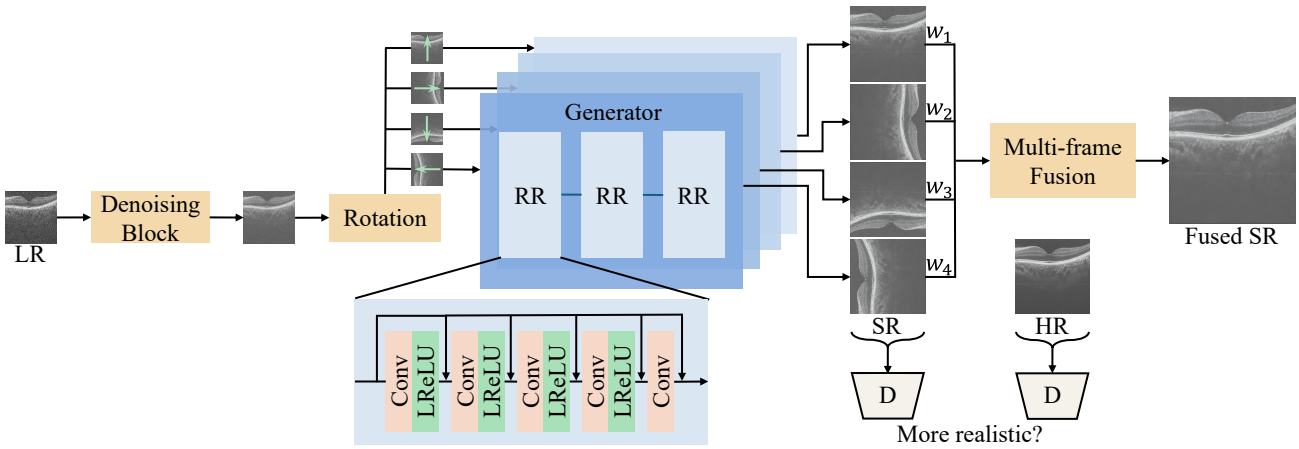


Figure 1: The network architecture of our MFGAN. The input image is denoised and rotated by multiples of 90 degrees before entering the four generator branches. The outputs are then rotated back and fused to produce a fused SR image. The block RR represents residual-in-residual blocks, and the block *D* represents the Discriminator.

2.2.2 Generative Adversarial Network

Inspired by previous works,[14,23,24] the Generator in our MFGAN contains two levels of residual layers, termed as residual-in-residual blocks (RR), to achieve dense connections. Following prior research,[25] we set a relativistic Discriminator to predict a real HR image x_{real} that is more realistic than a fake HR image x_{fake}:

$$\begin{aligned} D_r(x_{real}, x_{fake}) &= S(D(x_{real}) - E[D(x_{fake})]) \rightarrow 1, \\ D_r(x_{fake}, x_{real}) &= S(D(x_{fake}) - E[D(x_{real})]) \rightarrow 0, \end{aligned} \tag{1}$$

where $S(\cdot)$ denotes the sigmoid function, $D(\cdot)$ represents the original Discriminator output, and $E[\cdot]$ signifies the expectation operator. The Discriminator block *D* determines which of the SR and HR images is more realistic.

2.3 Multi-frame Fusion

As shown in *Figure 1*, we rotate each image by multiples of 90° and pass them through four parallel branches. After going through the RR blocks, the four output images are rotated back to their original orientations. We then take a weighted sum of the four images to produce the final output, with enhanced details. The multi-frame fusion process can be expressed as follows:

$$\hat{x} = \sum_{k=1}^{4} w_k \cdot x_{SR}^k, \tag{2}$$

where $\hat{x}$ represents the fused SR output, w_k is the weight for the k-th image, and x_{SR}^k refers to the SR output after rotation.

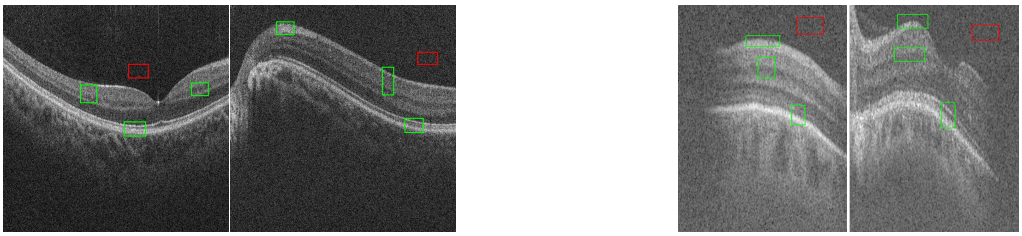
3. EXPERIMENTS

3.1 Data Preparation

We conducted experiments using two datasets, the PKU37 dataset[21] and the VIP Cup 2024 Test Set. The PKU37 dataset comprises 37 subjects, each with around 50 noisy OCT images at a resolution of 640 × 640 pixels. For each subject, a clean image was generated by averaging the frames of the noisy images. The VIP Cup 2024 Test Set includes data from 18 subjects, each with 70 to 300 B-scans at resolutions of either 300 × 150 or 300 × 200 pixels.

3.2 Implementation Details

We implemented the proposed method using the PyTorch framework and trained it with an NVIDIA GTX 4090 GPU. We adopted two commonly used evaluation metrics, *i.e.*, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). Following MPLN, we also measured two unsupervised metrics, *i.e.*, contrast-to-noise ratio (CNR) and mean-to-standard-deviation ratio (MSR) based on regions of interest (ROIs). Following the practice of TCFL,[21] we selected four ROIs to measure the non-reference metrics. As shown in *Figure 2*, three signal ROIs (green rectangles) are located at or near the retinal layers, and the background ROIs (red rectangles) were selected in the homogeneous region.



(a) PKU37 dataset *(b)* VIP Cup 2024 Test Set

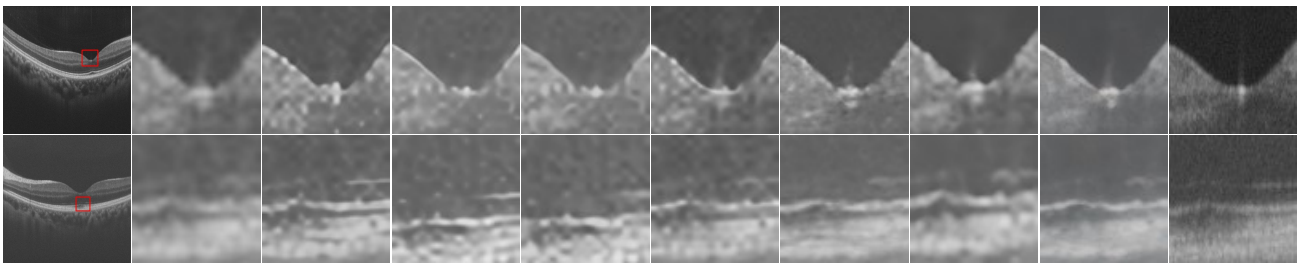
Figure 2: The ROIs selected for calculating CNR and MSR on the PKU37 dataset *(a)* and *(b)* the VIP Cup 2024 Test Set, respectively. The green rectangular ROIs are the foreground ROIs, and the red rectangular ROIs are the background ROIs.

Table 1: Quantitative comparison of various SR methods on the PKU37 dataset. The best and second-best results are highlighted in red and blue, respectively, and "-" indicates the result is not available.

Methods	×2				×4			
	PSNR↑	SSIM↑	CNR↑	MSR↑	PSNR↑	SSIM↑	CNR↑	MSR↑
EDVR	-	-	-	-	14.81	0.536	6.417	7.953
BasicVSR	-	-	-	-	12.96	0.510	6.464	7.988
SSL	-	-	-	-	12.91	0.506	6.344	7.705
EDSR	16.19	0.568	6.801	8.700	16.31	0.564	6.242	7.701
RealESRGAN	15.74	0.593	6.709	8.516	15.98	0.600	7.625	10.996
SwinIR	16.14	0.593	6.748	8.523	16.16	0.604	7.43	10.727
DAT	16.21	0.568	6.797	8.684	16.31	0.563	6.262	7.628
DKP	16.14	0.566	6.173	7.376	16.00	0.566	6.356	7.683
MFGAN (ours)	16.30	0.607	6.99	9.093	16.46	0.631	7.836	11.477

3.3 Experimental Results

We compared our method with other single image SR methods, including EDSR,[13] RealESRGAN,[20] SwinIR,[26] DAT[27] and DKP,[28] as well as video SR methods, including EDVR,[29] BasicVSR,[30] and SSL,[31] using OCT image sequences as inputs. We conducted experiments on the PKU37[21] dataset. The noisy images are denoised before being inputted to all the methods. *Table 1* presents the quantitative results of different methods, showing that our method outperforms both single-image and video SR methods. The visual results of different methods with a scale factor of 4 are shown in *Figure 3*. Video SR approaches[29–31] reconstruct shifted content and may produce incorrect information. EDSR[13] super-resolves the noises in the LR input. RealESRGAN[20] and DKP[26] deliver promising outputs, yet still maintain a certain level of blur. These results reveal that our method achieves superior performance in terms of visual quality and quantitative measures.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 3: Visual comparison of different methods on the PKU37 dataset by the scale of ×4. (a) Reference image, (b) LR of the area highlighted in red (denoised and enlarged by 4 times), (c) EDVR, (d) BasicVSR, (e) SSL, (f) EDSR, (g) RealESRGAN, (h) DKP, (i) MFGAN (ours), and (j) GT.

To further validate the robustness of our approach, we compared it with other methods on the VIP Cup 2024 Test Set, as shown in *Table 2* and *Figure 4*. Since reference images are not provided, we focused on CNR and MSR as non-reference

metrics. According to *Table 2*, our method obtained the best scores across CNR and MSR. As illustrated in *Figure 4*, these methods[13,26,27] produce similar results to the zoomed LR image, achieving little improvement. In contrast, our method achieves satisfactory output compared to others.

3.4 Ablation Studies

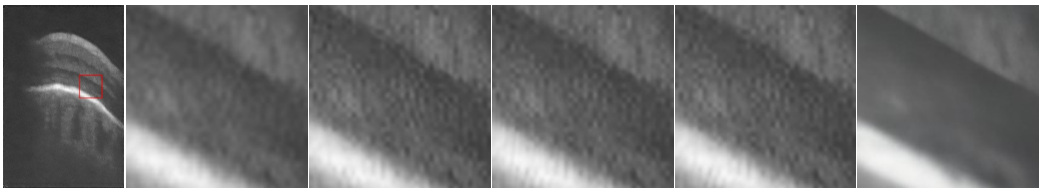
In this section, we explore our proposed method in two aspects: multi-frame fusion and denoising. The CNR and MSR scores of MFGAN on the VIP Cup 2024 dataset are illustrated as *Table 3*. It can be seen that both multi-frame fusion and denoising can improve the image quality.

Table 2: Quantitative comparison of different single image super-resolution methods on the VIP Cup 2024 Test Set. The best and second best results are highlighted in red and blue, respectively.

Scale	Metric	EDSR	SwinIR	DAT	DKP	**MFGAN**
×2	CNR↑	3.014	3.067	3.084	2.883	4.141
	MSR↑	7.395	7.411	7.457	8.394	12.206
×4	CNR↑	2.934	3.018	3.011	-	3.994
	MSR↑	7.265	7.261	7.347	-	13.777

Table 3: Quantitative results of MFGAN using proposed multi-frame fusion ("MF") and Denoising modules.

MF	Denoise	CNR	MSR
		1.756	8.230
✓		1.864	8.533
	✓	3.994	13.777
✓	✓	**4.263**	**15.532**



Reference LR EDSR[12] SwinIR[29] DAT[30] MFGAN (Ours)

Figure 4: Visual comparison of different methods on the VIP Cup 2024 Test Set with a scale factor of ×4.

4. CONCLUSION

In this paper, we have presented a novel framework for enhancing OCT retinal images that addresses the prevalent dual challenges of clinical OCT imaging, *i.e.*, noise and low resolution. Our approach integrates a denoising prior, a blind super-resolution module, and multi-frame fusion, collectively forming a robust pipeline that adapts effectively to varying degradation conditions in real-world retina OCT images. Unlike traditional methods, our method does not require paired high-resolution and low-resolution images, making it more suitable for clinical settings where obtaining such pairs is difficult. Experimental results on the PKU37[21] and VIP Cup 2024 datasets have validated the effectiveness of our method, demonstrating its superior performance in both visual quality and quantitative metrics. This method provides a promising advancement for the super-resolution and enhancement of retina OCT images.

5. DISCUSSION

The proposed method adopts an L1 loss function, which encourages pixel-wise accuracy. While this loss function performs well for overall structural restoration, it tends to over-smooth fine details. To address this limitation, future work can explore integrating more advanced loss function, *e.g.,* frequency domain perceptual loss, which can generate sharper textures and edges. Furthermore, while the current approach employs a blind degradation model, incorporating task-specific priors could make the method more robust to complex degradations observed in real-world OCT images.

REFERENCES

[1] Drexler, W. and Fujimoto, J. G., "State-of-the-art retinal optical coherence tomography," *Progress in retinal and eye research* **27**(1), 45–88 (2008).

[2] Gong, G., Zhang, H., and Yao, M., "Speckle noise reduction algorithm with total variation regularization in optical coherence tomography," *Optics express* **23**(19), 24699–24712 (2015).

[3] Young, M., Lebed, E., Jian, Y., and et al, "Real-time high-speed volumetric imaging using compressive sampling optical coherence tomography," *Biomedical optics express* **2**(9), 2690–2697 (2011).

[4] Chong, B. and Zhu, Y.-K., "Speckle reduction in optical coherence tomography images of human finger skin by wavelet modified bm3d filter," *Optics Communications* **291**, 461–469 (2013).

[5] Aum, J., Kim, J.-h., and Jeong, J., "Effective speckle noise suppression in OCT images using nonlocal means denoising filter with double gaussian anisotropic kernels," *Applied Optics* **54**(13), D43–D50 (2015).

[6] Zaki, F., Wang, Y., Su, H., Yuan, X., and Liu, X., "Noise adaptive wavelet thresholding for speckle noise removal in optical coherence tomography," *Biomedical optics express* **8**(5), 2720–2731 (2017).

[7] Xiao, J., Jia, W., and Lam, K.-M., "Feature redundancy mining: Deep light-weight image super-resolution model," in [*Proceedings of 2021 ICASSP*], 1620–1624, IEEE (2021).

[8] Xie, H., Huang, Z., Leung, F. H., Law, N., Ju, Y., Zheng, Y.-P., and Ling, S. H., "Satr: A structure-affinity attention-based transformer encoder for spine segmentation," in [*Proceedings of 2024 IEEE ISBI*], 1–5, IEEE (2024).

[9] Zuo, Y., Xiao, J., Chan, K.-C., Dong, R., Yang, C., He, Z., Xie, H., and Lam, K.-M., "Towards multi-view consistent style transfer with one-step diffusion via vision conditioning," *arXiv preprint arXiv:2411.10130* (2024).

[10] Chen, S., Chen, S., Guo, Z., and Zuo, Y., "Low-resolution palmprint image denoising by generative adversarial networks," *Neurocomputing* **358**, 275–284 (2019).

[11] Xu, M., Tang, C., Hao, F., Chen, M., and Lei, Z., "Texture preservation and speckle reduction in poor optical coherence tomography using the convolutional neural network," *Medical Image Analysis* **64**, 101727 (2020).

[12] Fang, L., Li, S., Nie, Q., Izatt, J. A., Toth, C. A., and Farsiu, S., "Sparsity based denoising of spectral domain optical coherence tomography images," *Biomedical optics express* **3**(5), 927–942 (2012).

[13] Lim, B., Son, S., Kim, H., Nah, S., and Mu Lee, K., "Enhanced deep residual networks for single image super-resolution," in [*Proceedings of 2017 CVPR workshops*], 136–144 (2017).

[14] Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C., "Esrgan: Enhanced super-resolution generative adversarial networks," in [*Proceedings of 2018 ECCV*], 0–0 (2018).

[15] Xiao, J., Ye, Q., Zhao, R., Lam, K.-M., and Wan, K., "Self-feature learning: An efficient deep lightweight network for image super-resolution," in [*Proceedings of the 29th ACM MM*], 4408–4416 (2021).

[16] Xiao, J., Jiang, X., Zheng, N., Yang, H., Yang, Y., Yang, Y., Li, D., and Lam, K.-M., "Online video super-resolution with convolutional kernel bypass grafts," *IEEE Transactions on Multimedia* **25**, 8972–8987 (2023).

[17] Xiao, J., Lyu, Z., Zhang, C., Ju, Y., Shui, C., and Lam, K.-M., "Towards progressive multi-frequency representation for image warping," in [*Proceedings of 2024 CVPR*], 2995–3004 (2024).

[18] Huang, Y., Li, S., Wang, L., Tan, T., et al., "Unfolding the alternating optimization for blind super resolution," *Advances in Neural Information Processing Systems* **33**, 5632–5643 (2020).

[19] Wang, L., Wang, Y., Dong, X., Xu, Q., Yang, J., An, W., and Guo, Y., "Unsupervised degradation representation learning for blind super-resolution," in [*Proceedings of 2021 CVPR*], 10581–10590 (2021).

[20] Wang, X., Xie, L., Dong, C., and Shan, Y., "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in [*Proceedings of 2021 ICCV*], 1905–1914 (2021).

[21] Geng, M., Meng, X., Zhu, L., Jiang, Z., Gao, M., Huang, Z., Qiu, B., Hu, Y., Zhang, Y., Ren, Q., et al., "Triplet cross-fusion learning for unpaired image denoising in optical coherence tomography," *IEEE Transactions on Medical Imaging* **41**(11), 3357–3372 (2022).

[22] Zhang, K., Liang, J., Van Gool, L., and Timofte, R., "Designing a practical degradation model for deep blind image super-resolution," in [*Proceedings of 2021 ICCV*], 4791–4800 (2021).

[23] Xiao, J., Ye, Q., Liu, T., Zhang, C., and Lam, K.-M., "Deep progressive feature aggregation network for multi-frame high dynamic range imaging," *Neurocomputing* **594**, 127804 (2024).

[24] Xie, H., Huang, Z., Leung, F. H., Ju, Y., Zheng, Y.-P., and Ling, S. H., "A structure-affinity dual attention-based network to segment spine for scoliosis assessment," in [*Proceedings of 2023 IEEE BIBM*], 1567–1574, IEEE (2023).

[25] Jolicoeur-Martineau, A., "The relativistic discriminator: a key element missing from standard gan," *arXiv preprint arXiv:1807.00734* (2018).

[26] Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R., "Swinir: Image restoration using swin transformer," in [*Proceedings of 2021 ICCV*], 1833–1844 (2021).

[27] Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., and Yu, F., "Dual aggregation transformer for image super-resolution," in [*Proceedings of 2023 ICCV*], (2023).

[28] Yang, Z., Xia, J., Li, S., Huang, X., Zhang, S., Liu, Z., Fu, Y., and Liu, Y., "A dynamic kernel prior model for unsupervised blind image super-resolution," in [*Proceedings of 2024 CVPR*], 26046–26056 (2024).

[29] Wang, X., Chan, K. C., Yu, K., Dong, C., and Change Loy, C., "Edvr: Video restoration with enhanced deformable convolutional networks," in [*Proceedings of 2019 CVPR workshops*], 0–0 (2019).

[30] Chan, K. C., Wang, X., Yu, K., Dong, C., and Loy, C. C., "Basicvsr: The search for essential components in video super-resolution and beyond," in [*Proceedings of 2021 CVPR*], 4947–4956 (2021).

[31] Xia, B., He, J., Zhang, Y., Wang, Y., Tian, Y., Yang, W., and Van Gool, L., "Structured sparsity learning for efficient video super-resolution," in [*Proceedings of 2023 CVPR*], 22638–22647 (2023).