DFA-Net: An efficient diffusion with frequency directional feature augmentation for low-light image enhancement Supplementary materials

Wengai Li, Zhaolin Xiao, Haonan Su

S.I. NETWORK STRUCTURE

A. DTCWT Decomposition and Reconstruction Process

As illustrated in Fig. S1, during the decomposition phase the input signal is recursively filtered and down-sampled by a factor of $\frac{1}{2^J}$ at level J, yielding a low-frequency sub-band and 6 high-frequency sub-bands at each level. The reconstruction phase subsequently restores the signal across scales via inverse filtering and up-sampling. Benefiting from its critically sampled and separable filter design, the DTCWT incurs low computational cost while suppressing down-sampling-induced oscillations and providing approximate shift invariance. Consequently, the DTCWT can capture directional features from high-frequency sub-bands at $\pm 15^\circ$, $\pm 45^\circ$, and $\pm 75^\circ$, providing an effective basis for directional degradation modeling in low-light scenarios. Moreover, the DTCWT and its inverse are implemented using fixed filter banks without introducing additional learnable parameters. This favorable trade-off between efficiency and directional expressiveness makes the DTCWT particularly suitable for lightweight low-light enhancement frameworks.

B. Details of diffusion refinement

Forward diffusion. During training, Gaussian noise is directly added to the clean high-frequency representation following the standard forward diffusion process:

$$\hat{\mathcal{F}}_t^h = \sqrt{\bar{\alpha}_t}\,\hat{\mathcal{F}}_0^h + \sqrt{1-\bar{\alpha}_t}\,\epsilon_t, \tag{S1}$$

where $\hat{\mathcal{F}}_0^h$ denotes the clean high-frequency representation, $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$, and $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Gaussian noise. **Reverse diffusion.** The reverse denoising process is parameterized by predicting the injected noise ϵ_t and computing the conditional mean accordingly:

$$\mu_\theta(\hat{\mathcal{F}}_t^h, M, t) = \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathcal{F}}_t^h - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(\hat{\mathcal{F}}_t^h, M, t)\right), \tag{S2}$$

where $\epsilon_\theta(\cdot)$ denotes the noise prediction network conditioned on the fused attention map M. During inference, we adopt DDIM sampler to progressively recover the high-frequency components from pure Gaussian noise, where the variance term is omitted and the reverse process is fully determined by the predicted mean. The noise schedule and diffusion timesteps are specified in the implementation details.

S.II. EXPERMENTS

A. Visualization Comparison on unpaired MEF, LIME, and NPE datasets

Fig. S2 presents a qualitative comparison on MEF[*S1*], LIME[*S2*], and NPE[*S3*] datasets between DFA-Net and current state-of-the-art methods. As shown, KinD[*S4*] and Zero-DCE[*S5*] fail to achieve sufficient brightness enhancement. Retinexformer[*S6*] introduces unavoidable noise, and lack of details. DiffLL[*S7*] and PyDiff[*S8*] produce relatively blurry results with degraded fine detail. SNR[*S9*] and URetinex-Net++[*S10*] suffer from noticeable contrast distortion and insufficient exposure control. Although HVI[*S11*] generates comparatively natural-looking results, it often introduces over-exposure artifacts, particularly in bright regions such as the sky, *e.g.* ., on the NPE dataset. In contrast, DFA-Net produces sharper details and more natural illumination across different datasets, further validating its cross-domain generalization capability.

B. Ablation on the position of FFAM in the diffusion refinement

Table SI presents an ablation study on the integration position of FFAM within the U-Net-based denoising network. Injecting the fused attention feature M at different upsampling stages leads to distinct trade-offs between restoration quality and computational cost. While introducing FFAM at all stages yields slightly higher PSNR, it incurs substantially increased parameters, FLOPs, and inference time, indicating considerable computational redundancy. In contrast, injecting FFAM only at the last upsampling stage achieves the highest SSIM and competitive PSNR with significantly reduced computational overhead. These results suggest that conditioning the reverse diffusion process with fused at the final upsampling stage is sufficient for effective structure preservation, thereby achieving an optimal balance between performance and efficiency. Consequently, this configuration is adopted in our final model.

This research has been funded in part by the National Natural Science Foundation of China (Grant No. 62371389, 62031023) and in part by the Scientific Research Program Funded by Shaanxi Provincial Education Department (Program No. 23JP105). (*Corresponding author: Zhaolin Xiao.*)

Wengai Li, Zhaolin Xiao, and Haonan Su are with the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China (e-mail: 1241211015@stu.xaut.edu.cn, xiaozhaolin@xaut.edu.cn, suhaonan@xaut.edu.cn).

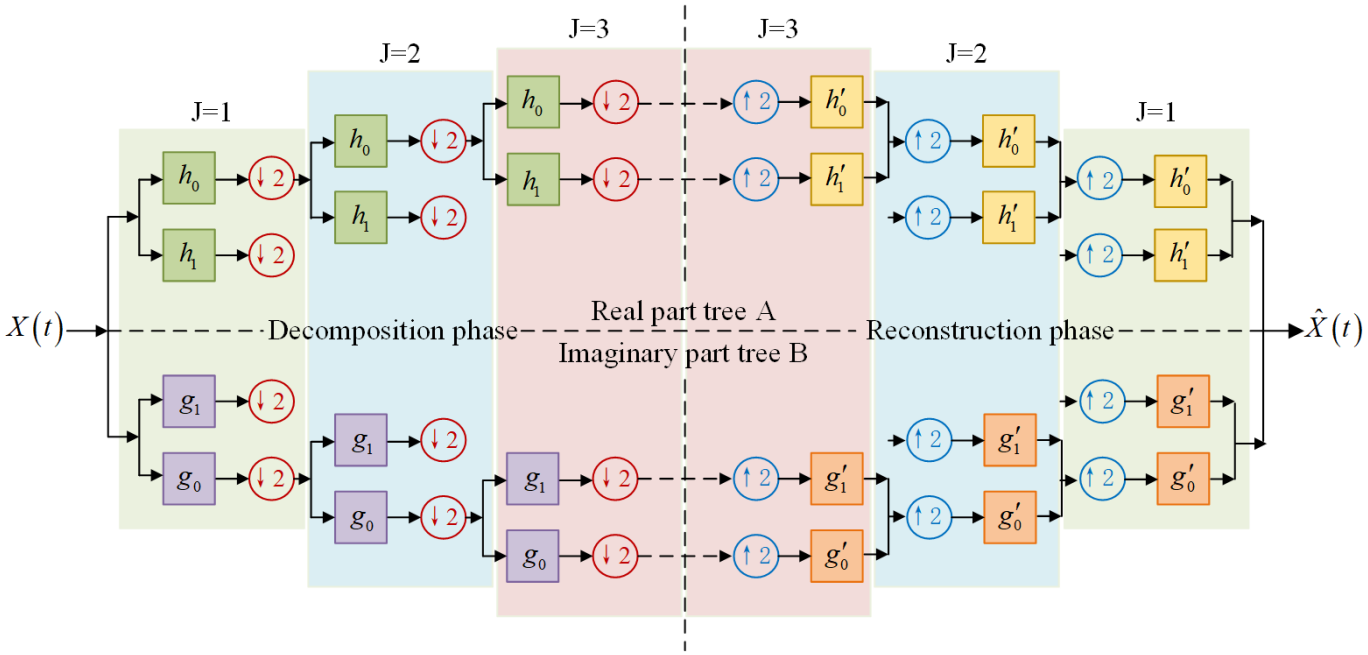

Fig. S1. The DTCWT decomposes the input signal $X(t)$ through two parallel filter trees, which apply paired filerter banks $\{h_0, h_1\}$ and $\{g_0, g_1\}$ to form the real and imaginary components of complex wavelet coefficients.

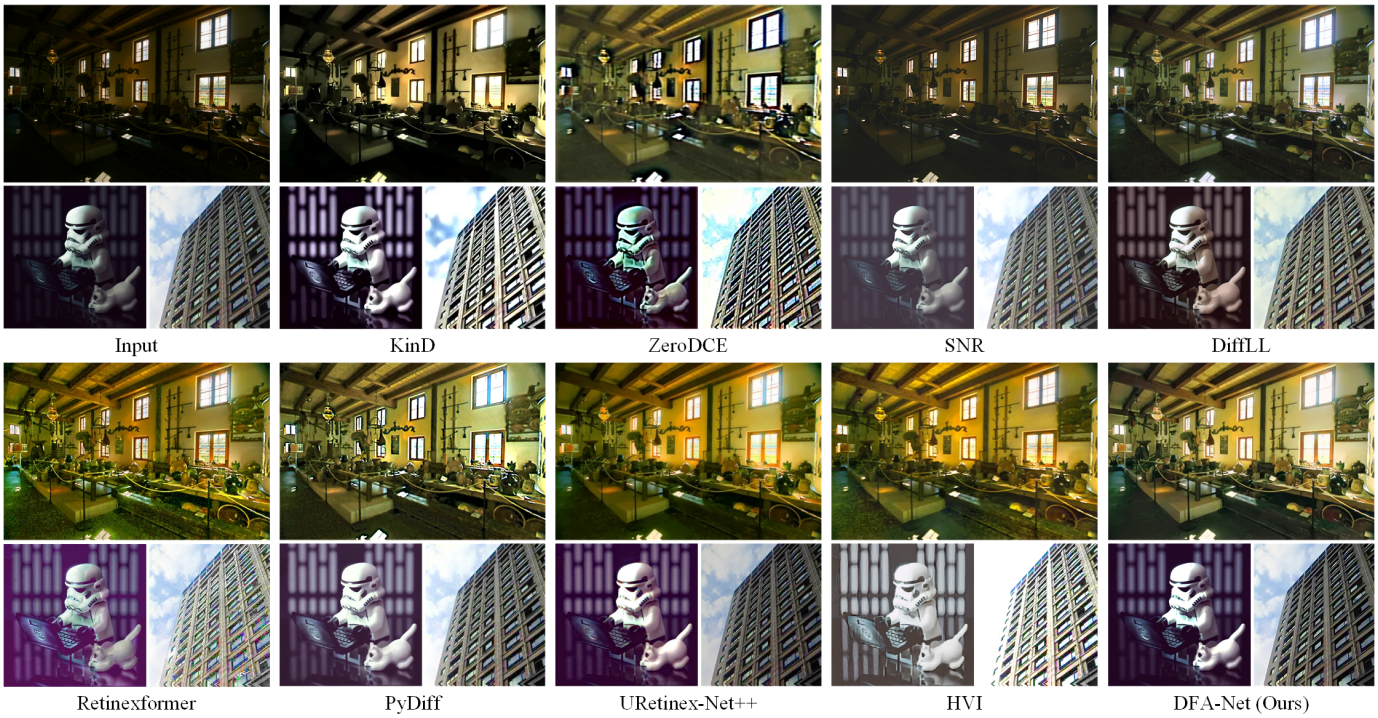

Fig. S2. Visual comparison of DFA-Net and competing methods on the unpaired MEF, LIME, and NPE datasets. The top row shows results on the MEF dataset, while the bottom row presents results on the LIME and NPE datasets.

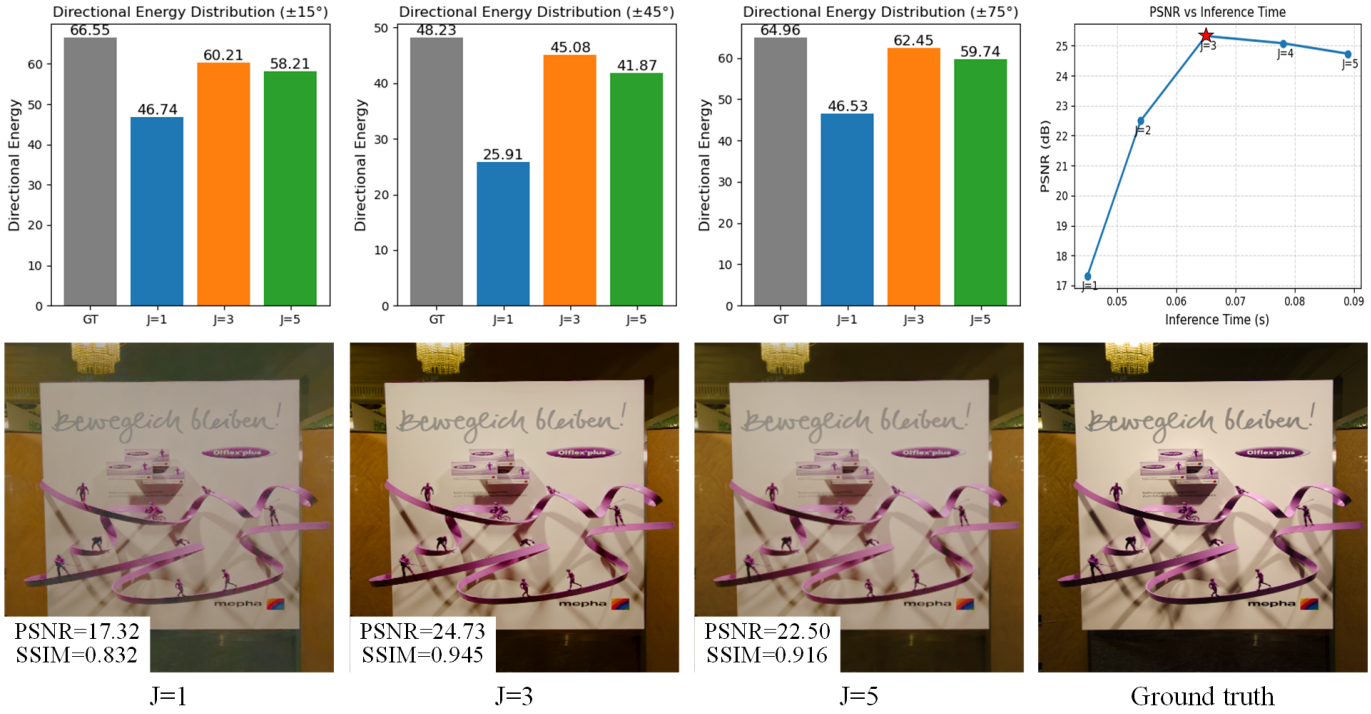

Fig. S3. The impact of different decomposition levels on the performance of DFA-Net.

TABLE SI
ABLATION STUDY ON THE INTEGRATION POSITION OF FFAM IN THE DIFFUSION DENOISING NETWORK.

FFAM Position	PSNR↑	SSIM↑	Params/M	FLOPs/G	Inference time/s
all stages	25.27	0.901	6.17	5.61	0.228
stage-1	24.96	0.905	4.94	2.38	0.058
stage-2	25.05	0.914	5.02	2.86	0.063
stage-3 (Ours)	25.23	0.931	5.26	3.37	0.065
-	24.81	0.905	4.77	2.13	0.057

TABLE SII
ABLATION STUDY ON THE NUMBER OF MDFA BLOCKS, MODEL DIMENSIONS, STATE DIMENSION L ON THE LOL-V1 AND LOLV2-REAL DATASET. THE OPTIMAL RESULTS OF EACH GROUP ARE MARKED IN RED.

Setting		LOL-v1		LOLv2-Real	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑
Number of blocks	1	24.1	0.844	21.33	0.857
	2	26.38	0.935	23.56	0.904
	3	26.33	0.893	24.03	0.891
Feature dimension	32	22.96	0.811	21.65	0.864
	64	23.51	0.851	23.65	0.887
	128	25.61	0.863	23.75	0.889
	256	24.69	0.84	23.19	0.878
State dimension L	4	23.01	0.842	23.45	0.803
	8	24.57	0.863	24.96	0.812
	16	25.84	0.945	25.08	0.855
	32	25.32	0.896	25.06	0.899
	64	25.11	0.867	25.31	0.876

C. The influence of the MDFA settings on the performance of DFA-Net

To identify the optimal model configuration, this paper performed ablation tests on the number of MDFA modules, feature dimensions, and state dimension L. The experimental results (in Table SII) demonstrate that stacking 2 MDFA blocks achieves the highest PSNR and SSIM on LOLv1, while yielding the best SSIM on LOLv2-Real with only a slight decrease in PSNR. Further ablation of feature dimension ablations indicated that a channel count of 128 provides the best trade-off between accuracy and efficiency on both datasets. Regarding the state dimension L, LOLv1 yields the highest metrics at $L = 16$. For the LOLv2-Real dataset, $L = 64$ corresponds to the highest PSNR, whereas $L = 32$ corresponds to the highest SSIM. Considering computational overhead and performance gains, experiments uniformly adopt $L = 16$.

D. The impact of decomposition levels in DTCWT on the performance of DFA-Net

Fig. S3 illustrates the impact of different decomposition levels on the performance of DFA-Net. In the top row, the first three columns compare the directional energy distributions at $\pm 15°$, $\pm 45°$, and $\pm 75°$ across different decomposition levels against the ground truth, highlighting directional features are preserved at each level. The fourth column reports the trade-off between restoration quality and efficiency by plotting PSNR against inference time for varying decomposition levels. The bottom row presents qualitative comparisons and the corresponding quantitative metrics for different decomposition levels. The results indicate that setting $J = 3$ yields the optimal visual quality and quantitative metrics. Moreover, the directional energy distributions at J=3 exhibit the greatest consistency with the ground truth. Importantly, this setting also achieves an effective balance between inference time and PSNR.

REFERENCES

[S1] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.

[S2] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 982–993, 2017.

[S3] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.

[S4] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1632–1640.

[S5] A. Mi, W. Luo, Y. Qiao, and Z. Huo, "Rethinking zero-dce for low-light image enhancement," *Neural Processing Letters*, vol. 56, no. 2, p. 93, 2024.

[S6] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *ICCV*, 2023, pp. 12 504–12 513.

[S7] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, "Low-light image enhancement with wavelet-based diffusion models," *ACM Trans. Graph.*, vol. 42, no. 6, Dec. 2023.

[S8] D. Zhou, Z. Yang, and Y. Yang, "Pyramid diffusion models for low-light image enhancement," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, ser. IJCAI '23, 2023.

[S9] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 714–17 724.

[S10] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Interpretable optimization-inspired unfolding network for low-light image enhancement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 4, pp. 2545–2562, 2025.

[S11] Q. Yan, Y. Feng, C. Zhang, G. Pang, K. Shi, P. Wu, W. Dong, J. Sun, and Y. Zhang, "Hvi: A new color space for low-light image enhancement," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5678–5687.