FTDB-Net: A Fourier Transform-Based Dual-Branch Low-Light Image Enhancement Network

Tianzhen Chen[1], Jie Liu[1], and Yi Ru[1] (✉)

[1] College of Computer Science, Inner Mongolia University, Hohhot 010021, IM, China
32309220@imu.edu.cn,yiru@imu.edu.cn

Abstract. This paper proposes a Fourier Transform-Based Dual-Branch Low-Light Image Enhancement Network, FTDB-Net, aimed at addressing issues such as color distortion, detail blurring, and noise in low-light conditions. By applying discrete Fourier transform, the image is divided into high-frequency and low-frequency regions. The high-frequency region contains rich details but is difficult to recover, relying on global information, while the low-frequency region is smoother and can be enhanced more easily using local information. FTDB-Net employs a dual-branch structure that combines the advantage of Transformers in capturing global information with the efficiency of convolutional networks in processing local information, optimizing the high-frequency and low-frequency regions separately. These two types of information are then combined through an adaptive fusion mechanism, significantly improving image clarity and detail. Experimental results show that FTDB-Net outperforms existing advanced methods on multiple benchmark datasets.

Keywords: Low-light Image Enhancement · Fourier Transform · Dual-branch structure ·Adaptive Feature Fusion.

1 Introduction

Low-light image enhancement holds significant importance in many real-world applications, for instance, after-dark monitoring, medical imaging, and astronomical observation. However, capturing and processing images in low-light conditions presents numerous challenges. Images captured under these conditions commonly exhibit reduced contrast, limited visibility, color distortion, and noise interference.

(a) Input (b) RtxFormer (c) ZeroIG (d) URtxNet (e) SNRNet (f) Ours

Fig. 1. Benchmark comparison with cutting-edge models (RtxFormer[1], ZeroIG[2], URetinexNet[3], SNRNet[4]) using the LIME dataset[5].

Conventional approaches for low-light image enhancement primarily employ methodologies including Retinex theory [6], histogram equalization, and gamma correction. While these methods can improve image quality to some extent, they often lead to issues like over-enhancement, noise amplification, and color distortion, with limited effectiveness in complex scenarios. Driven by the accelerated advancement of deep learning, significant progress has been made in the field of low-light enhancement, giving rise to various self-supervised and unsupervised methods.

Early approaches, such as RetinexNet [7], enhance images by decomposing them into reflectance and illumination components, but this often results in global color distortion. Unsupervised methods like Zero-DCE [9] use deep learning to estimate brightness adjustment curves, yet they exhibit noticeable noise under extreme low-light conditions. EnlightenGAN [10] employs generative adversarial networks for image enhancement but lacks effective noise suppression mechanisms. Most existing methods rely on CNNs, which struggle to capture distant spatial relationships. To address this, researchers have turned to Transformer architectures, such as RtxFormer [1] and SNR-Net [4]. While these methods improve global feature modeling, they still suffer from uneven brightness and artifacts.

To advance low-light image enhancement, we introduce FTDB-Net, a dual-branch Fourier transform-based network. The algorithm begins with frequency-domain decomposition, separating high-frequency (detail) components from low-frequency (illumination) regions: high-frequency elements are enhanced through global feature restoration, while low-frequency components are refined using local information. Our architecture processes distinct frequency bands independently via its dual-path design before adaptive feature fusion.Experimental results confirm substantial improvements in image clarity and detail preservation (Fig. 1). Key innovations include:

1) We present FTDB-Net, a novel low-light enhancement framework employing discrete Fourier transform for spectral decomposition into high- and low-frequency components, each processed through dedicated pathways.

2)The network uses a dual-branch structure that combines the global information capture of Transformers with the local feature extraction of CNNs and employs an adaptive fusion mechanism to preserve details lost during downsampling, improving image quality and details.

3) Empirical validation across five reference datasets indicates our model's significant improvements compared to contemporary best-performing methods, verified through both statistical analysis and visual inspection.

2 Method

2.1 Overview of the Model Framework

The overall framework is shown in Fig. 2, where low-light images sequentially pass through the two branches of the network: Detail Enhancement Multi-Scale Network, DEMS-Net and Fourier Transform Attention Network, FTA-Net. This setup enablesadaptive learning of different enhancement operations for high-frequency and

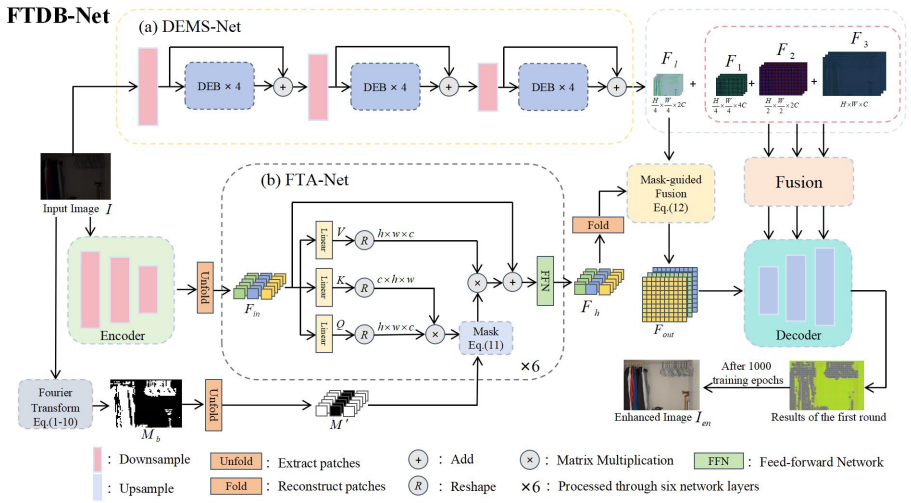


Fig. 2. The overall structure of FTDB-Net.

low-frequency regions. In FTA-Net, the low-frequency and high frequency regions of the input image I are first computed and marked with 0 and 1, respectively, resulting in the mask M_b. The mask M_b is then adjusted using the "Unfold" operation, which transforms the input feature map into column vectors representing local regions, to obtain M'. The Transformer in FTA-Net utilizes the input feature map F_{in} and M' to guide the attention mechanism and generate the high-frequency feature map F_h.

In DEMS-Net, the low-frequency feature map F_l, as well as the first-stage feature map F_1, second-stage feature map F_2, and third-stage feature map F_3 are obtained after processing through the depthwise enhancement block, DEB [12]. Subsequently, F_h and F_l are fused using the mask for guidance. Additionally, we use Fusion [12] to integrate the upsampled components from FTA-Net with the feature maps F_1, F_2 and F_3, enhancing image details and producing the result of the first round. After 1000 rounds of training, the final enhanced image I_{en} is obtained.

2.2 Mask Calculation

FTA-Net employs Fourier-transform-based dynamic masking (Fig.3) to automatically separate high-frequency details from low-frequency structures. The mask guides the Transformer to focus on high-frequency regions while preserving original low-frequency information through feature fusion.

Image Blocking and Frequency Domain Transformation

To calculate the mask, the input image I is divided into non-overlapping blocks I_b of size $B \times B$, where the block indices (i, j) traverse all possible positions in the image.

$$i \in \{0, B, 2B, \ldots, H - B\}, j \in \{0, B, 2B, \ldots, W - B\}, \quad (1)$$

Here, $H \times W$ represents the image size, and i and j denote the starting row and column

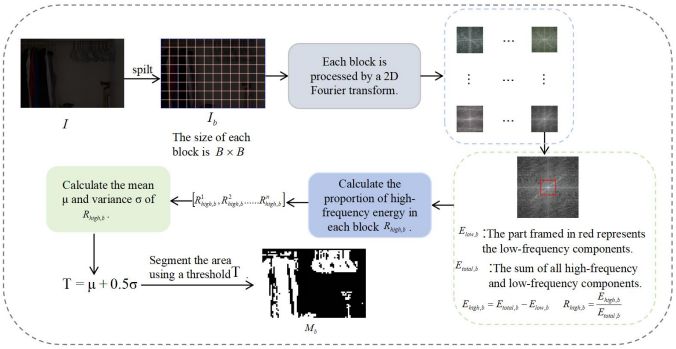


Fig. 3. The process of calculating the mask M_b.

Here, $H \times W$ represents the image size, and i and j denote the starting row and column coordinates of the block, respectively. For each block I_b, a 2D Discrete Fourier Transform (DFT) is applied, as shown in (2):

$$F_b(u,v) = F_{2D}(I_b), \tag{2}$$

Via spectrum reorganization with zero-frequency at the center, we obtain $F_{shift,b}$, and then compute the magnitude spectrum,(3):

$$M_b(u,v) = |F_{shift,b}(u,v)|, \tag{3}$$

The center of the frequency spectrum, denoted by (c_r, c_c), is defined as the geometric center of the block size, i.e., $c_r = \lfloor B/2 \rfloor$ and $c_c = \lfloor B/2 \rfloor$.

The calculation of the high-frequency energy ratio

Low-frequency Region Definition: Based on the spectral center(c_r, c_c), a square region with side length $2r$ is defined, where $r = \lfloor B/8 \rfloor$. The empirical basis for choosing $r = \lfloor B/8 \rfloor$ is that significant B low-frequency energy in natural images is typically concentrated within a range of approximately $B/4 \times B/4$ around the spectral center (which accounts for 6.25% of the total frequency domain). This ratio can effectively isolate high-frequency components while avoiding excessive removal of low-frequency information.

Energy Calculation: The low-frequency energy $E_{low,b}$ and the total energy $E_{total,b}$ are respec tively denoted by Equations (4) and (5):

$$E_{low,b} = \sum_{u=c_r-r}^{c_r+r} \sum_{v=c_c-r}^{c_c+r} M_b(u,v), \tag{4}$$

$$E_{total,b} = \sum_{u=0}^{B-1} \sum_{v=0}^{B-1} M_b(u,v), \tag{5}$$

To quantify the proportion of high-frequency components within an image for spectral component discrimination, we calculate the high-frequency energy and its proportion, as shown in (6) and (7):

$$E_{high,b} = E_{total,b} - E_{low,b}, \tag{6}$$

$$R_{high,b} = \frac{E_{high,b}}{E_{total,b}}, \tag{7}$$

Calculate the mean and variance of the high-frequency energy ratios [$R^1_{high,b}$, $R^2_{high,b}$, ... , $R^N_{high,b}$], with N representing the complete set of blocks.

The calculation of the high-frequency energy ratio

Dynamic Threshold Determination: To distinguish between high-frequency and low-frequency regions in an image, We initially compute an adaptive threshold using first-order statistics ($\mu \pm \sigma$) of high-frequency component ratios across all spatial blocks.We start by computing $R_{high,b}$ for all blocks and then calculate the mean μ and standard deviation σ, as shown in (8):

$$\mu = \frac{1}{N}\sum_{b=1}^{N} R_{high,b}, \quad \sigma = \sqrt{\frac{1}{N-1}\sum_{b=1}^{N}\left(R_{high,b} - \mu\right)^2}, \tag{8}$$

with N representing the complete set of blocks. The threshold calculated in this way will be adjusted according to the characteristics of different images. The final dynamic threshold is:

$$T = \mu + 0.5\sigma, \tag{9}$$

As shown in (9), T denotes the threshold.

Mask Generation

For each block I_b, a binary mask is generated based on the threshold T:

$$M_b = \begin{cases} 1, & if\ R_{high,b} > T \\ 0, & otherwise. \end{cases}, \tag{10}$$

If the $R_{high,b}$ of a block exceeds the threshold T, the mask value is set to 1; otherwise, it is set to 0. After processing all N blocks, the corresponding mask M_b is generated, as shown in Fig.3, where the value of each $B \times B$ region is consistent with M_b. In compliance with the Transformer's input constraints, the mask M_b is adjusted to M'. This mask guides the network to enhance high frequency details while preserving low-frequency structures.Fig. 4 shows the mask generated by calculating the threshold through the Fourier transform, as well as the result after the first round of network processing using the mask M_b.

2.3 Mask-guided self-attention

Using the mask, we can accurately identify the high-frequency and low-frequency

regions of the image. Next, we will explain in detail how the mask functions within the attention mechanism, as shown in the following equation:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}} + (1 - M')\varepsilon\right)V, \quad (11)$$

In the attention mechanism, the query Q, key K, value V are the three core components. By calculating the similarity score matrix between Q and K, we can reflect the relevance of each query to all keys. The value d_k represents the dimension of the keys, used to scale the attention scores and prevent excessively large values. Additionally, the mask-matrix M' indicates the validity of each position. The value of 1 indicates that the posi-tion is valid, while the value of 0 indicates it is invalid. The value of ε is set to -1×10^9, approximating negative infinity. When $M' = 1$, the position is valid, and its information is processed; when $M' = 0$, the position is invalid, and its information is ignored. Ultimately, the model calculates the attention weights for different regions and multiplies them by the value V , ensuring that the output includes information only from valid positions.

Following the generation of spectral feature maps (F_h, F_l) in Fig. 2, we perform weighted fusion using attention mask M_b to yield the enhanced output F_{out} through the following operation:

$$F_{out} = (1 - M_b) \times F_l + M_b \times F_h, \quad (12)$$

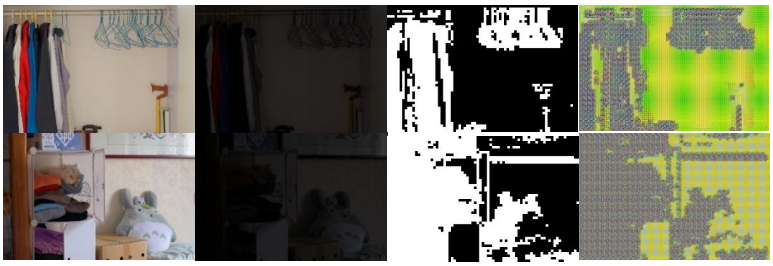
The multiplication symbol represents the weighted combination of the low-frequency feature map F_l and high-frequency feature map F_h using the mask M_b.

2.4 Feature Extraction and Fusion

To effectively process low-frequency information, we designed DEMS-Net based on the DEB and Fusion modules from [12]. The network generates multi-scale feature maps (F_1-F_3) and integrates them through upsampling fusion. As shown in Fig.5, DEB employs a serial architecture of depthwise separable (DEConv) and standard convolutions to enhance image details and textures. The fusion module incorporates CGA attention [12], implementing feature optimization via three attention mechanisms: spatial attention combines average/max pooling for weight generation; pixel attention emphasizes key details; channel attention performs adaptive fusion using weights w and $1 - w$.

2.5 Loss Functions

The proposed network employs a triple-loss optimization framework (Eqs. 13-16): (i) an L_1 loss ensures pixel-wise accuracy by minimizing absolute errors between enhanced (I_{en}) and reference (I) images; (ii) a perceptual loss L_{vgg} enhances visual fidelity through deep feature-space alignment; and (iii) a chromatic consistency loss L_{color} maintains color accuracy across the CIELAB color space.



(a) (b) (c) (d)

Fig. 4. Dynamic thresholding for segmenting high-frequency and low-frequency regions in an image. Figure (a) shows the original image, (b) shows the dark image, (c) shows the mask Mb, which is calculated from the dark image in (b), and (d) shows the result after the first round of network processing, where high-frequency and low-frequency regions are processed separately.

$$L_r = \sqrt{\|I_{en} - I\|_2 + \epsilon^2}, \tag{13}$$

$$L_{vgg} = \| \phi(I_{en} - I) \|_1, \tag{14}$$

$$L_{color} = \sum_k \angle((I_{en})_k, (I)_k), \tag{15}$$

$$L_{total} = \lambda_1 L_r + \lambda_2 L_{vgg} + \lambda_3 L_{color}, \tag{16}$$

In (13), L_r represents the L_1 loss, with parameter ϵ set to 1×10^{-3} for computational stability. In (14), L_{vgg} denotes the perceptual loss, where ϕ refers to the operation of extracting features from a pre-trained VGG network, and the L_1 loss is used to compute the VGG feature distance between I and I_{en}. In (15), L_{color} represents the color loss, where k stands for each pixel in the image. The color loss treats the RGB values as vectors and computes the cosine similarity between the two vectors to determine the loss. In (16), L_{total} is the total loss, with λ_1, λ_2, and λ_3 as weight factors set to 1, 0.1, and 1, respectively, to balance the contributions of different losses.

3 EXPERIMENTS

3.1 Datasets and Implementation Details

We conducted tests using multiple publicly available low-light image datasets, includeing five datasets: LOL, LIME, DICM, MEF and NPE. To quantitatively evaluate the enhancement effects, we employed Peak Signal-to-Noise Ratio (PSNR)and Structural Similarity Index Measure (SSIM) as evaluation metrics on the LOL dataset, while the last four datasets were assessed using the noreference image quality assessment metric, Natural Image Quality Evaluator (NIQE).

FTDB-Net was developed using PyTorch framework on an NVIDIA RTX 3060 GPU. We adopted the Adam optimizer (β_1=0.9, β_2=0.999) with initial learning rate

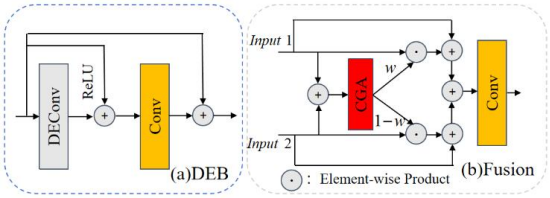


Fig. 5. Structural diagram of the DEB depthwise enhancement block and Fusion module.

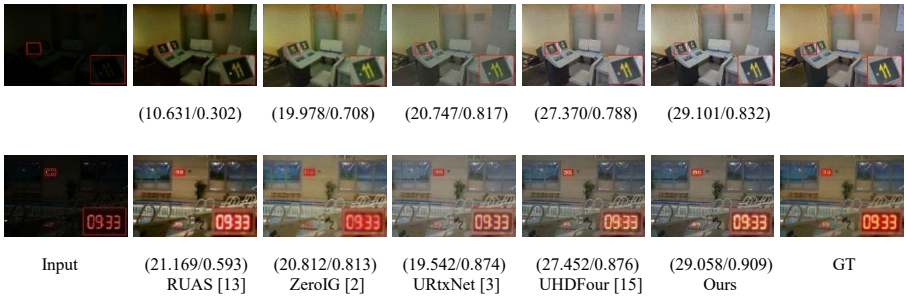


Fig. 6. Visual results on the LOL dataset using different methods, with PSNR(dB)/SSIM values in parentheses.

10^{-4}, dynamically adjusted via LambdaLR scheduler. Training involved 1,000 iterations on original image pairs to minimize reconstruction error, requiring approximately 3 days on the LOL dataset.

3.2 Comparative Experiments

Results on the LOL dataset. We compared our method with several state-of-the-art algorithms. As shown in Fig. 6, our method achieves the best performance in color and texture restoration, closely matching the ground truth. In contrast, ZeroIG, RUAS and URtxNet suffer from significant color distortion, while UHDFour generates noticeable artifacts.Moreover, the quantitative results in Table 1 show that our method achieves the highest PSNR and ranks second in SSIM, highlighting its strength in noise reduction and structural restoration. Additionally, our model has 39.48M parameters and only 6.17G FLOPs, demonstrating higher computational efficiency compared to other methods. Results on unpaired datasets. We tested our on the DICM, LIME, MEF and NPE datasets and compared it with RAUS, RetinexNet, SNRNet, URtxNet, RtxFormer and ZeroIG. Using NIQE as the evaluation metric, the results in Table 2 show that our model outperforms all others across the four datasets. Fig. 7 demonstrates our method's superior image quality, highlighting its significant advantage in image restoration tasks.

Fig. 7. Visual results of different methods on multiple datasets (NIQE values in parentheses): the first rows are from the MEF [16] dataset, the second row is from the DICM [7] dataset, and the last row is from the NPE [17] dataset.

Table 1. Results comparison for the LOL dataset and LOL-v2-Syn dataset, with bold indicating the best and underlined indicating the second best.

Methods	LOL [7]		LOL-v2-Syn [8]		Complexity	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM	Para(M)	FLOPs(G)
RetinexNet [7]	16.77	0.43	17.13	0.80	0.84	587.47
Zero-DCE [9]	14.86	0.56	16.60	0.78	0.08	1.30
RUAS [13]	16.41	0.50	16.55	0.65	0.003	0.83
EnGAN [10]	17.61	0.65	16.57	0.73	114.35	61.01
URtxNet [3]	19.84	0.82	18.75	0.83	0.36	56.93
FourLLIE [14]	21.91	0.77	**24.65**	**0.92**	0.12	2.55
UHDFour [15]	23.09	0.82	23.64	0.90	17.54	57.42
ZeroIG [2]	22.18	0.77	15.77	0.75	0.12	8.10
RetinexFour [11]	<u>23.25</u>	**0.89**	22.73	<u>0.91</u>	1.38	68.45
Our	**23.43**	<u>0.84</u>	<u>24.43</u>	**0.92**	39.48	6.17

3.3 Ablation Study

Our ablation study systematically evaluates three critical components: the attention mask M', fusion module, and DEMS-Net feature extractor (denoted as γ_1, γ_2, γ_3).The baseline configuration retains only FTA-Net after component removal. Quantitative evaluation on the LOL dataset (Table 3) using PSNR/SSIM metrics reveals: (1) performance degradation occurs with any component removal, (2) the mask M' (γ_1) demonstrates particularly strong impact on PSNR (Δ>2.1dB), and (3) comprehensive validation of each module's contribution to the overall architecture.

Table 2. Comparison of results for DICM, LIME, MEF and NPE datasets using the NIQE metric, with bold indicating the best and underline indicating the second best.

Methods	LIME [5]	DICM [7]	MEF [16]	NPE [17]
	NIQE			
RetinexNet [7]	4.598	4.479	4.416	4.594
RUAS [13]	4.225	4.781	3.828	5.678
SNRNet [4]	5.919	6.133	4.035	5.451
URtxNet [3]	4.352	3.798	3.789	4.692
RtxFormer [1]	4.279	3.291	<u>3.458</u>	<u>4.007</u>
ZeroIG [2]	<u>4.133</u>	<u>3.203</u>	3.896	4.321
Our	**3.994**	**2.954**	**3.171**	**3.933**

Table 3. Ablation studies on LOL dataset. The bold text indicates that our complete FTDB-Net performs the best.

Baseline	*γ1*	*γ2*	*γ3*	PSNR(dB)	SSIM
✓	×	×	×	22.643	0.836
✓	×	✓	✓	22.363	0.832
✓	✓	×	✓	23.111	0.830
✓	✓	✓	×	22.989	0.832
✓	✓	✓	✓	23.430	0.841

4 CONCLUSION

We propose FTDB-Net, a Fourier-based image enhancement network that decomposes images via DFT into high/low-frequency components. These components are processed separately: high frequencies by FTA-Net (Transformer encoder with Fourier-guided attention mask M') and low frequencies by DEMS-Net (hybrid depthwise/standard CNNs), then adaptively fused. Our method outperforms state-of-the-art approaches across multiple low-light enhancement benchmarks.

Acknowledgement. This work was supported by Inner Mongolia University High Level Talent Research Launch Project under Grant 10000-21311201/068 and the fund of supporting the reform and development of local universities (Disciplinary construction).

References

1. 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12504–12513, 2023.

2. Y. Shi, D. Liu, L. Zhang, Y. Tian, X. Xia, and X. Fu, “ZERO-IG: Zero-Shot Illumination-Guided Joint Denoising and Adaptive Enhancement for Low-Light Images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3015–3024, 2024.
3. W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, “Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5901–5910, 2022.
4. 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*, pp. 17714–17724, 2022.
5. 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, 2016.
6. E. H. Land, “The retinex theory of color vision,” *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.
7. C. Wei, W. Wang, W. Yang, and J. Liu, “Deep retinex decomposition for low-light enhancement,” *arXiv preprint arXiv:1808.04560*, 2018.
8. W. Yang, W. Wang, H. Huang, et al., “Sparse Gradient Regularized Deep Retinex Network for Robust Low-light Image Enhancement,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2072–2086, 2021.
9. C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, “Zero-reference deep curve estimation for low-light image enhancement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1780–1789, 2020.
10. Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, “Enlightengan: Deep light enhancement without paired supervision,” *IEEE Transactions on Image Processing*, vol. 30, pp. 2340–2349, 2021.
11. H. Li, J. Wang, H. Yang, X. Tang, and F. Xu, “Learning Semantic-aware Retinex Network with Spatial Frequency Interaction for Low-light Image Enhancement,” in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2024.
12. Z. Chen, Z. He, and Z.-M. Lu, “DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention,” *IEEE Transactions on Image Processing*, 2024.
13. R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, “Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10561–10570, 2021.
14. C. Wang, H. Wu, and Z. Jin, “Fourllie: Boosting low-light image enhancement by Fourier frequency information,” in *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 7459–7469, 2023.
15. C. Li, C.-L. Guo, M. Zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, “Embedding Fourier for Ultra-High-Definition Low-Light Image Enhancement,” in *ICLR*, 2023.
16. 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.
17. S. Wang, J. Zheng, H.-M. Hu, and B. Li, “Naturalness preserved enhancement algorithm for nonuniform illumination images,” *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3538–3548, 2013.