IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 5 July 2024, accepted 22 July 2024, date of publication 29 July 2024, date of current version 9 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3434531

RESEARCH ARTICLE

WMANet: Wavelet-Based Multi-Scale Attention Network for Low-Light Image Enhancement

YANGJUN XIANG[1], GENGSHENG HU[2], MEI CHEN[1], AND MAHMOUD EMAM[2,3]
[1]School of Media and Design, Hangzhou Dianzi University, Hangzhou 310018, China
[2]Shangyu Institute of Science and Engineering Company Ltd., Hangzhou Dianzi University, Shaoxing 312300, China
[3]Faculty of Artificial Intelligence, Menoufia University, Shebeen El-Kom 32511, Egypt

Corresponding author: Gengsheng Hu (40675@hdu.edu.cn)

ABSTRACT Low-light images captured at night often suffer from improper exposure, color distortion, and noise, which degrades the image quality and have a negative influence on subsequent applications. Many existing deep learning-based methods enhance low-light images through spatial domain, which may sacrifice the original image information. In this paper, we put forward a deep learning network for enhancing low-light images based on wavelet transform. We utilize the wavelet transform to divide the image into various frequency scales and then analyze the frequency characteristics of different low-light images in the wavelet domain. The proposed network comprises a low-frequency restoration subnet and high-frequency reconstruction subnet that uses an optimal coefficient of wavelet decomposition to construct a frequency pyramid. Furthermore, we utilized different attention mechanisms to extract frequency information from different images, gradually restoring the brightness information and details of low-light images. Additionally, we utilized a self-constructed multi-scale exposure low-light image dataset for training. Numerous experiments on publicly accessible datasets and our established dataset show that the proposed approach quantitatively and qualitatively surpasses state-of-the-art approaches, particularly for real and complex low-light scenarios. Furthermore, our method produces better visual effects than others and performs well in real-time and real-word downstream vision tasks.

INDEX TERMS Low-light image enhancement, wavelet transform, multi-scale, attention, deep learning.

I. INTRODUCTION

The rapid development of mobile devices has allowed people to take photos and share them on social media anytime and anywhere. However, acquiring high-quality images in low-light situations is a tough task. Increasing ISO, aperture, long exposure, and flash can improve the first shot image. A high ISO enhances the sensitivity of the sensor while amplifying noise. Furthermore, many devices do not have large-aperture lenses and large-sized sensors. Moreover, prolonged exposure can cause motion blurring and limited usage scenarios. The use of flash lights up the subject of the shot but makes the photo visually unpleasant. On the other hand, most users do not know the equivalent exposure operation and take up a lot of time in post-processing the photos. They only wanted to obtain satisfactory photos quickly. As a result, low-light image enhancement (LLIE) has always been an important issue of concern to the community, and developing an efficient LLIE method is essential for improving the viewing experience and subsequent usability of images.

Deep learning-based approaches have delivered impressive results in low-level visual tasks, far surpassing traditional image enhancement methods, such as defogging [1] and high dynamic range reconstruction [2]. Traditional approaches like histogram equalization [3], [4] and gamma correction [5], [6], stretch the image contrast to restore photos. The Retinex theory [7], which is founded on the color perception of the human eyes and color invariance, determines the reflective properties of objects by eliminating the influence of lighting. Image fusion [8] combines different images captured from the same scene to synthesize a single high-quality image to restore luminance information. Furthermore, Cooley et al. [9], decomposed an image into amplitude and phase spectra

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.

using the Fast Fourier Transform method, which was then processed separately and inverted in the spatial domain to recover the image. However, these handcrafted constraints and priors must be sufficiently adaptive, and the enhanced results may result in strong noise or excessive enhancement. Learning-based methods usually sample paired low-light/normal-light datasets for training and use regularization terms to constrain the loss. Thus, the model can learn the mapping relationship between the paired images. There is no unified optimal lighting, and each user has a different preference for image lighting. Therefore, simply mapping low-light images to specific lighting levels is inappropriate.

This paper suggests a deep wavelet transform-based multi-scale LLIE model to partially solve the above problems inspired by deep Fourier exposure correction [10] and Multi-level Wavelet-CNN [11]. First, we constructed a multi-scale exposure image dataset to make learning more efficient and infer rich illumination adjustments. Our model comprises a Low-frequency Repair Network (LRN) module and a High-frequency Reconstruction Network (HRN) module. First, the input images were converted to the wavelet domain, and then the frequency pyramids were constructed using the optimal wavelet decomposition coefficients. Following the principle of the divide-and-conquer method, LRN is used to recover the luminance information for the low-frequency portion, in which the Simplified Attention (SA) mechanism is designed to reasonably assign weights to different lighting regions, enhance the interaction between regions with different exposure levels, and remove the transposed convolution to avoid the artifacts in the image, HRN to reconstruct the details and reduce the noise, where the feature extraction block (FEB) is applied to efficiently extractor the features of each high-frequency portion, and then the cross-attention mechanism is utilized to fuse the different high-frequency information. The loss function is carefully crafted to constrain the reconstruction and colors such the low-light images can be effectively recovered with reasonable exposure, precise details, and vivid colors. The following are the main contributions of this paper:

- We propose a low-light image enhancement network that employs wavelet transform to construct a frequency pyramid, extracting multi-scale critical features from the training dataset for efficient image enhancement.
- We further designed a low-frequency restoration network (LRN) and a high-frequency reconstruction network (HRN), where the SA of the LRN efficiently assigns weights to different light regions, and the HRN utilizes FEB and cross-attention for efficient learning and feature fusion.
- A new dataset was constructed that contained seven scales of exposure levels and 22,500 images, each with expert-retouched references.
- We executed comprehensive experiments on publicly available datasets and our established dataset, to evaluate the robustness and effectiveness of the put forward model.

II. RELATED WORK

This section briefly reviews various existing LLIE methods, which can be categorized into two main types: traditional-based and deep-learning-based LLIE approaches. Traditional-based enhancement approaches include histogram equalization, image fusion, and Retinex theory. Lately, deep-learning-based approaches have become the most popular.

A. TRADITIONAL-BASED LLIE METHOD

Histogram Equalization (HE) [3], [12], [13] is a classic method for image enhancement. If an image has an even distribution of pixel values among all grayscale levels, it exhibits high contrast and broad dynamic range. Consequently, the HE algorithm adjusts the grayscale using a uniform probability density function to reproduce dark area details. However, simply changing the mapping relationship of pixels in HE needs to be more flexible, it may ignore the structural information of the image, which results in insufficient enhancement and noise amplification issues.

To address the limited dynamic range problem in images and the loss of structural information, a stack-based high dynamic range (HDR) [14] method was proposed to combine multi-exposure images into an HDR map and subsequently compress the dynamic range of the map using a tone-mapping operator thereby reproducing scenes in the dark. Unlike the HDR method, the multi-exposure fusion (MEF) method [15], [16] generates HDR images by using multiple exposures of the same scene and captures the best details of each image based on the exposure time. Furthermore, Merianos and Mitianoudis [17] used two image fusion methods, one for brightness and the other for color. These methods yield good results but require several images of the same scene. Consequently, achieving fast real-time image enhancement is complex, and images with low global illumination exhibit poor restoration effects.

Retinex theory [7] typically decomposes images into reflectance and illuminance components, assuming that the reflectance component is concord under any illumination condition. Therefore, the LLIE was formulated as an illumination estimation problem. Wang et al. [18] presented a method for maintaining natural details when processing non-uniform illumination images. Additionally, Guo et al. [19] refined the luminance map by obtaining the peak intensity at each pixel location and incorporating structured priors. Retinex theory-based methods have apparent strengths when it comes to color image enhancement. However, using Gaussian convolutional kernels for illumination estimation cannot preserve edges [20], which may lead to halos in certain areas with clear boundaries or overexposed images.

B. DEEP LEARNING- BASED LLIE METHOD

Recent years, the popularity of deep learning has driven the development of the LLIE field, and many excellent techniques have emerged [21], [22], [23], [24], [25], [26],

[27], [28], [29], [30]. Wei et al. [24] presented Retinex-Net by combines the retinex theory and neural networks to estimate luminance maps and restore low-light images. Furthermore, Sobbahi and Tekli [25] proposed LLHF-Net to incorporate homomorphic filtering into a neural network to perform image-to-frequency learning, and designed extended models that can be used in different deep learning architectures. The EnlightenGAN presented by Jiang et al. [26] is an unsupervised generative adversarial network (GAN) that uses multiple-degree discriminators, self-regularization perception loss, and attention mechanisms to constrain non-paired data learning. Additionally, Li et al. [27] put forward a Zero-Reference Deep Curve Estimation (Zero-DCE) method that accepts low-light images as input and computes higher-order curves to improve the image by adjusting its dynamic range. Ma et al. [28] designed a self-calibrated illuminant system (SCI) that utilized a cascade illuminant learning methodology with weights sharing. Moreover, the SNR-Net proposed by Xu et al. [29] combines CNN and Transformer to improve low-light images by performing long-range and short-range operators on different regions of the image using the signal-to-noise ratio prior of the images. Cai et al. [30] put forward the Retinexformer, which introduced the Transformer into the Retinex model by first calculating the illumination information and then utilizing the Illumination-Guided Transformer to non-locally interact with different lighting regions of the image. Moreover, Huang et al. [10] introduced (FECNet), by converting the image to the frequency domain via Fourier Transform and then applied an amplitude sub-network and phase sub-network approach to adjust the image exposure.

Overall, deep learning-based approaches typically process images in the spatial domain, the structure of the Transformer requires larger computational resources and guidance from image priori, and they rarely explore potential solutions from the frequency domain. However, the global information in an image can be effectively captured in the frequency domain. Some studies [31], [32], [33] have investigated strategies to improve neural networks' generalization ability and data enhancement using Fourier transform. Generally, the Fourier transform divides an image into low and high-frequency parts, which may not capture the local structural information of the image, leading to further information leakage after the inverse transformation. Most of the current methods use paired datasets that contain only low-light and the corresponding ground-truth images, and further ignore multiple exposure levels in between. Consequently, methods trained on these datasets have poor generalization in real complex low-light scenes and are prone to color distortion. We will describe our proposed methodology and dataset in Sections III and IV, respectively, to address the above issues.

III. PROPOSED METHOD

A. MOTIVATION

In this paper, we introduce a new wavelet-based perspective for LLIE, that facilitates the restoration of frequency domain information. Following the literature [34] the wavelet transform is an expression between the Fourier and spatial domains that decomposes a signal into a series of independent, spatially orientated frequency channels. The low-frequency component comprises most of the luminance information, whereas the high-frequency component comprises the structural and textural information. First, let's relive the operations and properties of the wavelet transform, given a low-light image $x \in \mathbb{R}^{H\times W\times C}$, we down-sample and transform the input image into one low-frequency portion and three high-frequency portions using a two-dimensional discrete wavelet transform (2D-DWT) with a Symlet wavelet function, which can be formulated as follows:

$$\{A_{low},\ V_{low},\ H_{low},\ D_{low}\} = \text{2D - DWT(x)}, \tag{1}$$

where A_{low} denotes the low-frequency coefficient of the image, which impacts the global brightness of the image, and V_{low}, H_{low}, D_{low} denote the high-frequency coefficients of the image, which represent the details of the image in the horizontal, vertical, and diagonal directions, respectively.

As seen in Fig. 1, the low-light image can be restored to approximately the same light level as normal-light by exchanging the low-frequency coefficients and then inverting them using a two-dimensional inverse discrete wavelet transform (2D-IDWT). Therefore, the key to restoring low-light images in the wavelet domain is to restore the low-frequency coefficients of the low-light images, i.e., restore the low-frequency coefficients that are consistent with the normal-light images. It can also be seen in Fig. 1 that detail is lost in low-light images compared to normal-light images, and is not sufficient enough to restore the low-frequency coefficients but the high-frequency coefficients, which represent the detail information, also need to be enhanced. Therefore, we designed a low-frequency restoration network (LRN) and high-frequency reconstruction network (HRN) to achieve light restoration and detail reproduction from low-light images.

B. OVERALL NETWORK FRAMEWORK

Fig. 2 illustrates the general framework of the suggested approach. First, the low-light image is converted into wavelet domain using 2D-DWT. To construct the wavelet pyramid structure and coarse-to-fine extraction of image features, we decomposed the image l times and obtained low-frequency and high-frequency coefficients each time. This can be formulated as follows:

$$\left\{{A^l}_{low},\ {V^l}_{low},\ {H^l}_{low},\ {D^l}_{low}\right\} = \text{2D - DWT}({A^{l-1}}_{low}), \tag{2}$$

$$\left[C^l,\ L^l\right] = \text{2D - DWT}(x,\ l,\ `sym2'), \tag{3}$$

where ${A^l}_{low},\ {V^l}_{low},\ {H^l}_{low},\ {D^l}_{low} \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times C}$, $l\in[1,\ l]$ and C^l represents the low-frequency coefficients after each decomposition corresponding to the ${A^l}_{low}$. The L^l represents the three high-frequency coefficients after

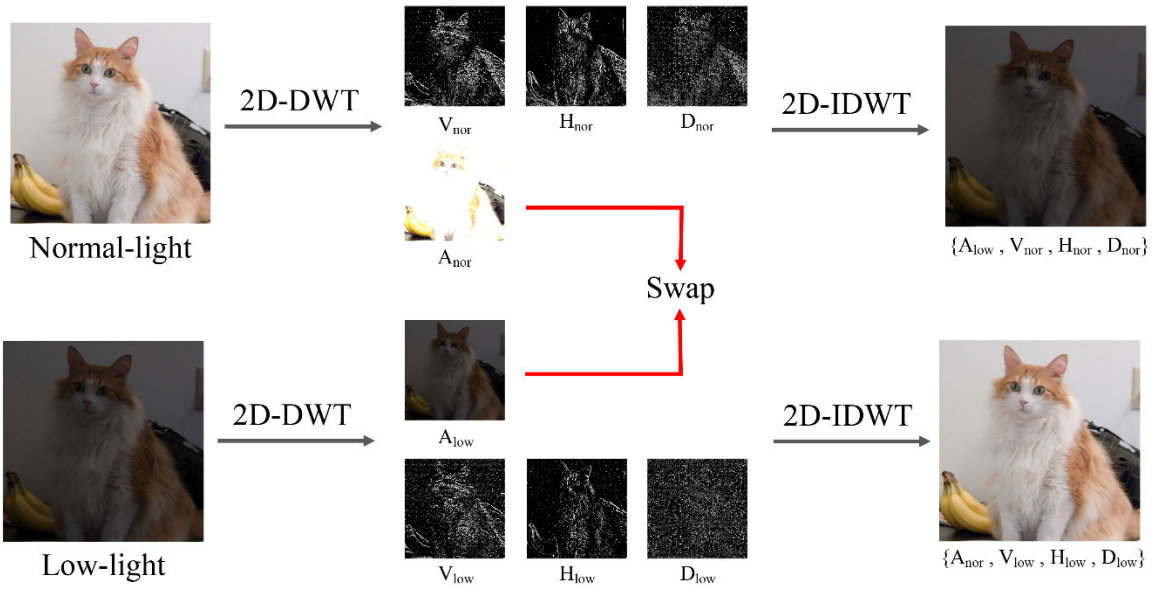

FIGURE 1. Schematic diagram of low-frequency wavelet coefficient exchange between a low-light image and a normal-light image.

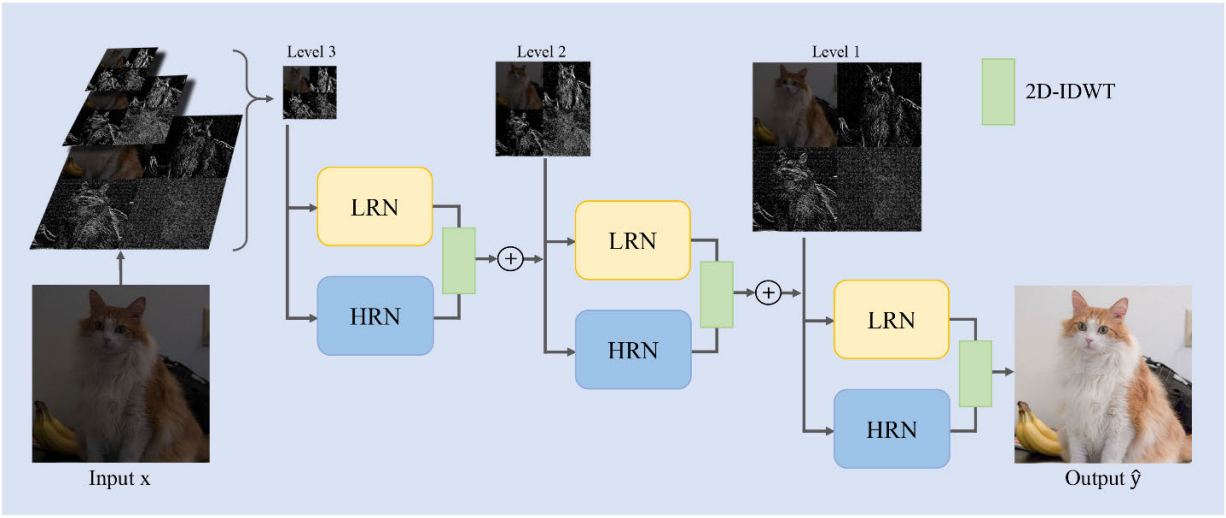

FIGURE 2. The overall framework of the put forward approach, consisting of LRN and HRN.

each decomposition corresponding to $\{V_{low}, H_{low}, D_{low}\}$, and sym2 is a wavelet function with approximate symmetry proposed by Daubechies [35], to reduce the phase distortion during the decomposition and reconstruction of images. Subsequently, we input the low-frequency coefficients C of each layer of the wavelet pyramid into the low-frequency restoration network and the high-frequency coefficients L into the high-frequency reconstruction network. The enhanced low-frequency and high-frequency portions are reconstructed by the 2D-IDWT to rebuild the recovered low-light image $\widehat{y}$, which is formulated as follows:

$$\widehat{y}^{l-1} = \text{2D - IDWT}(\left\{\widehat{A}^{l}_{low}, \widehat{V}^{l}_{low}, \widehat{H}^{l}_{low}, \widehat{D}^{l}_{low}\right\}). \quad (4)$$

Such a refinement transformation process continues until the final output image is generated.

C. LOW-FREQUENCY RESTORATION NETWORK (LRN)

As shown in Fig. 3, the LRN comprises three encoder modules (E1, E2, E3) and two decoder modules (D1, D2), which receive the low-frequency portion of the extracted wavelet pyramid decomposition $A^{l}{}_{low}$. Each encoder block consists of the following sequence of two convolutional layers and the LReLU [36] activation layer, which are downsampled using max-pooling. Finally, critical information is extracted by a designed simple attention (SA) mechanism before moving to the next level of encoder. In the input multi-scale low-light image, some regions are underexposed, while others are normally exposed. The attention mechanism can assign different weights to different regions of illumination and enhance non-local interactions between different exposure levels, resulting in a natural light distribution in the enhanced image. The same structure

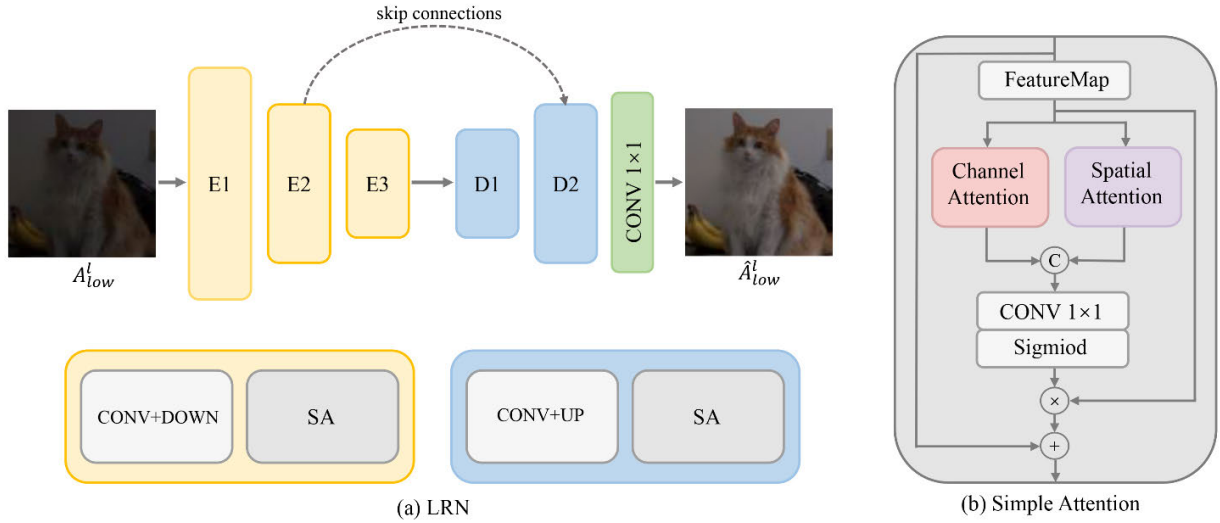

FIGURE 3. The main structures of: (a) LRN module. (b) A simple attention.

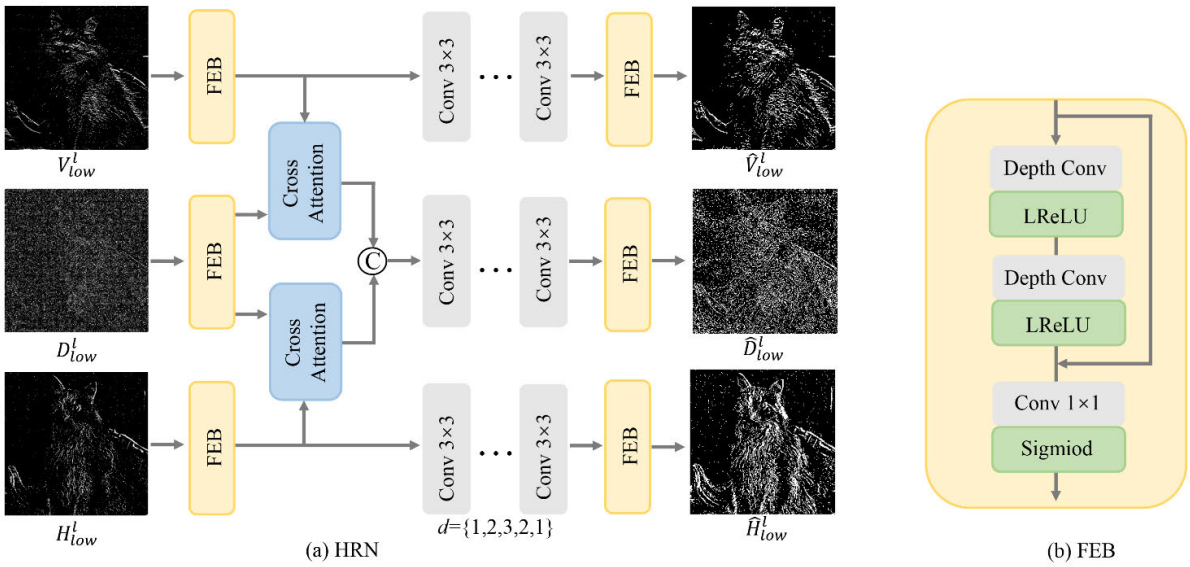

FIGURE 4. The main structures of: (a) HRN module. (b) Feature Extraction Block (FEB).

was used for decoder blocks D1 and D2. Considering that the frequent use of transpose convolution at multiple levels produces noticeable checkerboard artifacts [37], instead of using transpose convolution for upsampling as in U-Net [38], we employ line nearest neighbor upsampling and then convolution instead. The final LRN uses a fully connected layer to synthesize the extracted features to obtain the final output $\widehat{A}^l_{low}$.

D. HIGH-FREQUENCY RESTORATION NETWORK (HRN)

To enhance low-light images and restore rich details, we propose a high-frequency reconstruction network for enhancing high-frequency information. As shown in Fig. 4, we first extract the features from high-frequency coefficients V^l_{low}, H^l_{low}, and D^l_{low} using three feature extraction blocks (FEB), which mainly consist of depth-separable convolution [39] and LReLU, which can effectively reduce the computational and parametric quantities, and also increase the inference and operation efficacy of the proposed model. The diagonal details are then supplemented using two cross-attention layers [40] by utilizing the horizontal and vertical details. Subsequently, we proposed a progressive dilation module that extracts local information in the first and last convolutions while improving the sensory field to utilize distant information better in the middle convolutions. The gradual increase and decrease in the dilation rate can efficiently prevent the grid effect of the image. Eventually, we used the previous three FEB to obtain the reconstructed high-frequency coefficients $\widehat{V}^l_{low}$, $\widehat{H}^l_{low}$, and $\widehat{D}^l_{low}$. After obtaining the recovered low-frequency coefficients and reconstructed high-frequency coefficients, we can apply 2D-IDWT to obtain the output $\widehat{y}^{l-1}$ with a scale of l-1.

E. LOSS FUNCTION

The put forward model was trained in an end-to-end manner with the objective of minimizing a loss function:

$$L_{total} = L_{low} + L_{high}, \tag{5}$$

where L_{low} denotes the loss in the low-frequency restoration network and L_{high} denotes the loss in the high-frequency network.

1) LOW FREQUENCY LOSS

The low-frequency loss comprises the absolute loss and structural similarity SSIM loss [41], minimizing the discrepancy between the restored low-frequency coefficients and those of the normal-light image. The specific formulae are as follows:

$$L_{low} = \sum_{l=1}^{l} \left|\widehat{C}^{l}_{low} - \widehat{C}^{l}_{nor}\right|_1 + (1 - SSIM\ (\widehat{C}^{l}_{low},\ \widehat{C}^{l}_{nor})), \tag{6}$$

where $\widehat{C}^{l}_{low}$ denotes the low-frequency coefficients of the lth layer of the wavelet decomposition of the network output, and $\widehat{C}^{l}_{nor}$ is the low-frequency coefficients of the lth layer of the wavelet decomposition of the corresponding normal-light image.

2) HIGH FREQUENCY LOSS

We employed the Mean Square Error (MSE) loss to measure the variation among details. Additionally, we add Total Variation (TV) loss [42] to ensure result smoothness and hence prevent over-enhancement, avoiding artifacts and amplifying noise. This can be formulated as follows:

$$L_{high} = \lambda_1 \sum_{l=1}^{l} \|\widehat{L}^{l}_{low} - \widehat{L}^{l}_{nor}\|_2 + \lambda_2 \sum_{l=1}^{l} TV(\widehat{L}^{l}_{low}), \tag{7}$$

where $\widehat{L}^{l}_{low}$ represents the high-frequency coefficients of the lth layer of the wavelet decomposition of the network output, and $\widehat{L}^{l}_{nor}$ represents the high-frequency coefficients of the lth layer of the wavelet decomposition of the corresponding normal-light image. λ_1 and λ_2 are the weights of each term. In the experiments, we set λ_1 and λ_2 to 0.01 and 0.1, respectively.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section first presents implementation details, test datasets, and evaluation metrics. Then, quantitative and qualitative results are introduced, and the results are contrasted with the current state-of-the-art low-light image enhancement approaches. Finally, an ablation experiment is conducted to show the efficacy of each component of the put forward model.

A. IMPLEMENTATION DETAILS

We built our network on Pytorch and trained and tested it on a core i9 3.7GHz CPU and two NVIDIA RTX 3060 GPUs. The entire network was optimized using the Adam optimizer [43], with an initial learning rate of 0.0001 and a cosine annealing schedule [44]. The images were randomly cropped into 512 × 512 segments for training, and the corresponding batch size was set to 64 according to the GPU memory. The wavelet transform ratio l was set to 3.

B. DATASET

1) OUR ESTABLISHED DATASET

Unlike existing low-light datasets, we created a novel vast collection of images with diverse exposure levels. Fig. 5 shows the differences in exposure levels across our established dataset and the LOL dataset [24]. Compared with our dataset, the LOL dataset had a relatively limited exposure level of coverage. We need many training images to train our proposed model correctly, including realistic multi-scale exposure levels and the corresponding ground truth. Our dataset came from the MIT-Adobe FiveK dataset [45], with 5,000 raw-RGB images, and the related sRGB images were manually rendered by five professional photographers. For every raw image, we utilized Adobe Lightroom Classic to accurately simulate the nonlinear camera rendering process. We rendered every original image with varying digital exposure values (EV) for realistic multi-scale exposure. In detail, we utilized relative EVs -4.0, -3.0, -2.0, and +1.0, to render underexposed and overexposed images. The ground truth, as the correct exposure of our dataset, was an image that was manually modified by a professional photographer.

(a) Our established dataset (b) LOL dataset

FIGURE 5. Comparison of the exposure levels of our dataset with those of the LOL dataset.

We meticulously chose 4,500 images out of the initial 5,000 raw images and eliminated silhouette photos, art photos, and images with exposure errors and noticeable noise. Our established dataset comprises a total of 22,500 8-bit sRGB images with disparate digital exposure settings. The dataset was bifurcated into three categories: (i) a training set with 17,500 images and (ii) a testing set with 5,000 images. Each set presented distinct scenarios for training and testing.

2) TEST DATASET

Besides the dataset we built in the previous subsection, we also used the paired dataset LOL [24], which encompasses 500 pairs of natural low-light/normal-light photography. Furthermore, we employed the following unpaired datasets: LIME [19] (with 10 images), DICM [46] (with 64 images), MEF [47] (with 17 images), VV [48] (with 24 images), and ExDark [49] (with 7358 images) to assess the capability of the suggested model.

TABLE 1. Performance evaluation comparison of various low-light image enhancement approaches on our established dataset using four evaluation metrics and runtime. The bold-red font indicates the best result, and the second-best result is formatted with bold-blue font.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DeltaE ↓	RT ↓
Retinex-Net [24]	15.4830	0.7494	0.2042	19.3032	1.0853s
LLHF-Net [25]	15.6805	0.6539	0.3527	21.1113	0.5713s
EnlightenGAN [26]	14.6498	0.8386	0.0911	19.7605	1.4443s
Zero-DCE++ [27]	20.5625	0.8005	**0.0756**	9.5131	0.3957s
SCI [28]	16.1944	0.8794	0.0867	14.4930	0.4066s
URetinex-Net [53]	17.8801	0.8896	0.0781	15.1268	0.5493s
UHDFour [54]	**25.8506**	**0.9032**	0.1124	**5.6399**	0.6449s
SNR-Net [29]	13.7354	0.7026	0.2693	23.5748	**0.0723s**
Retinexformer [30]	13.9804	0.7687	0.1135	22.7905	0.1223s
Ours	**30.0850**	**0.9437**	**0.0252**	**2.8494**	**0.0562s**

TABLE 2. Performance evaluation comparison of various low-light image enhancement approaches on the paired LOL dataset using four evaluation metrics and runtime. The bold-red font indicates the best result, and the second-best result is formatted with bold-blue font.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DeltaE ↓	RT ↓
Retinex-Net [24]	17.6157	0.4199	0.3670	17.1965	0.4933s
LLHF-Net [25]	18.8734	0.6917	0.2377	12.6498	0.2597s
EnlightenGAN [26]	19.2420	0.7211	0.1699	15.9915	0.5968s
Zero-DCE++ [27]	19.1402	0.5296	0.1781	17.4070	0.1799s
SCI [28]	17.2628	0.6232	0.1702	27.1878	0.1848s
URetinex-Net [53]	16.9918	**0.8484**	**0.0696**	17.6969	0.2065s
UHDFour [54]	17.6390	0.7210	0.3936	14.5683	0.2424s
SNR-Net [29]	21.3770	0.8132	0.1359	13.6107	**0.0268s**
Retinexformer [30]	**24.1951**	0.8206	0.0932	**8.4632**	0.0556s
Ours	**23.8892**	**0.8366**	**0.0547**	**6.9615**	**0.0273s**

3) EVALUATION METRICS

For the paired dataset, the following four criteria metrics were used to assess the pixel precision and perceived quality of the output results: (i) peak signal-to-noise ratio (PSNR), (ii) structural similarity (SSIM), (iii) Learned perceptual image patch similarity [50] (LPIPS), and (iv) Delta-E [51] (ΔE). Furthermore, we used the Natural Image Quality Evaluator [52] (NIQE) as a non-reference assessment metric for the non-paired datasets. This reconstruction metric does unnecessary ground truth images and is more consistent with human visual habits. The higher PSNR and SSIM values denote better image quality, whereas lower LPIPS and ΔE values denote better image quality. Furthermore, we provide runtime analysis of each method to further evaluate its computational efficiency and real-time performance.

C. COMPARISION WITH STATE-OF-THE-ART

We compared our method with nine state-of-the-art deep learning-based method-ologies, including Retinex-Net [24], LLHF-Net [25], EnlightenGAN [26], Zero-DCE++ [27], SCI [28], URetinex-Net [53], UHDFour [54], SNR-Net [29] and Retinexformer [30]. For a fair comparison, we employed the openly implemented and recommended parameter settings provided by the corresponding authors of each approach to enhance low-light images.

1) QUANTITATIVE COMPARISONS

We first assessed the efficacy of diverse methods on our established dataset (as detailed in Section IV-B1). The results achieved by each technique are summarized in Table 1. The put forward approach exceeds other state-of-the-art approaches on all four indicators, has a significant lead, and has the shortest running time.

In order to ascertain the model's generalization capacity, we further assessed the LOL dataset without retraining or fine-tuning. Basically, LOL paired dataset is the training dataset used by most of the existing methods. The performance evaluation comparison for these experimental results is presented in Table 2. Our suggested method lags slightly behind Retinexformer and URetinex-Net in two metrics of pixel accuracy. In the perceptual metrics LPIPS and ΔE are both best, indicating that our method can have better visual quality. It is also only behind SNR-Net in runtime. Finally, there was no reference evaluation indicator NIQE. Due to the lack of authentic reference images, our proposed multi-scale wavelet model exerts the corresponding advantages. As presented in Table 3, our approach takes the lead for multiple datasets and provides the best results on average.

2) QUALITATIVE COMPARISONS

To verify the effectiveness of the suggested model, a series of comparison experiments were performed for qualitative

TABLE 3. Summary of natural image quality evaluator (NIQE) scores on various low-light image datasets, with red being the best result and blue the second best.

Method	Dataset					Avg. ↓
	LIME	MEF	VV	DICM	ExDark	
Retinex-Net [24]	2.6660	3.0447	2.4943	3.5595	4.4795	3.2488
LLHF-Net [25]	3.3493	3.2279	5.3634	3.3360	5.9247	4.2403
EnlightenGAN [26]	**2.0141**	2.5231	4.2326	2.5028	4.4226	3.1390
Zero-DCE++ [27]	2.5732	**2.4522**	2.5714	**1.9468**	**2.9146**	**2.4916**
SCI [28]	2.5516	2.4661	**1.9252**	3.5884	3.1552	2.7373
URetinex-Net [53]	3.0039	3.0765	**1.9254**	2.9321	3.5677	2.9011
UHDFour [54]	4.2382	3.4628	3.6318	3.1523	4.2435	3.7457
SNR-Net [29]	3.2315	2.7782	6.2230	3.1033	6.3586	4.3389
Retinexformer [30]	2.8139	3.0750	3.4055	2.3868	4.5976	3.2558
Ours	**2.2810**	**2.3153**	1.9412	**1.7173**	**3.1209**	**2.2751**

(a) Input (b) Retinex-Net (c) LLHF-Net (d) EnlightenGAN
(e) Zero-DCE++ (f) SCI (g) URetinex-Net (h) UHDFour
(i) SNR-Net (j) Retinexformer (k) **Ours** (l) Ground truth

FIGURE 6. Visual comparison results with various approaches on a complex multi-light source image from our established dataset, zooming in to get the best viewing effect.

verification. Figures 6 and 7 visually compare two challenging cases, using a complex multi-light source image from our established dataset (Figure 6 (a)) and an overall low-light image with almost no feature (Figure 7 (a)). From these two figures, it can be seen that our put forward approach can restore more details and better contrast while keeping the overall exposure more reasonable. From these two figures, it can be seen that our put forward approach can restore more details and better contrast while keeping the overall exposure more reasonable. Other methods, such as Retinex-Net and LLHF-Net, greatly amplify noise, produce artifacts, and unnaturally enhance results. SCI and SNR-Net did not improve the image significantly and were less effective than others. URetinex-Net and UHDFour recover well, but the image suffers from blurring and color undersaturation.

In addition, we provide a visual comparison of the ExDark dataset without reference images, as illustrated in Fig. 8. Our suggested approach achieves suppression of noise and artifacts by restoring people's exposure without over-enhancing the image. Overall, the proposed approach produces better visual effects than the other approaches, which further proves its excellent generalization ability for different types of low-light images.

3) LOW-LIGHT OBJECT DETECTION

This section examines the effects of different LLIE approaches as preprocessing on the efficacy of object detection under low-light conditions. Specifically, we made a comparison with the ExDark dataset, which has a composition of 7358 images obtained in actual nighttime

FIGURE 7. Visual comparison results with various approaches on a very low-light image with from the paired LOL dataset, zooming in to get the best viewing effect.

FIGURE 8. Visual comparison results with various approaches on the ExDark dataset, note that ExDark dataset does not have reference image, zooming in to get the best viewing effect.

settings, including 12 classes of objects. Our proposed method and the comparison method are first applied as a preprocessing step, and then the final results are evaluated using the latest object detection technique Yolov10 [55]. Different enhancement methods impact the performance of the object detection algorithms, as shown in Fig. 9. The LLHF-Net approach blurs the image and creates artifacts resulting in poor detection. The Zero-DCE approach detects more objects but the noise problem makes the targets less credible. Our method can detect the most number of object

FIGURE 9. Visual comparison of object detection results before and after enhanced by various approaches on the ExDark dataset, zooming in to get the best viewing effect.

classes and the highest confidence level while improving the image quality.

D. ABLATION STUDY

In this section, we execute two sets of experiments to evaluate the impact of employing various parts in the proposed model. We trained all of them using the implementation details presented in Section IV-A and test them on our constructed dataset.

1) WAVELET DECOMPOSITION SCALE

Wavelet decomposition takes progressively finer spatial or frequency-domain steps for high frequencies so that it can be focused on analyzing arbitrary details of the image, but too many wavelet decomposition layers do not mean that the final enhancement result will be better, so it is essential to identify the optimum number of wavelet decomposition layers. We selected sym2 wavelet and used five different decomposition scales for the experiment, and the results are illustrated in Table 4. It can be noticed that, as the wavelet scale l gradually increases, all four metrics become higher, runtime is also increasing. However, when $l = 4$ or 5, the evaluation metrics were not as ideal as when $l = 3$. In light of the trade-off between computational efficiency and performance, it was determined that the wavelet decomposition scale l should be set to 3.

2) KEY MODULES

We previously introduced the simplified attention mechanism (SA) and feature extraction block (FEB) in the low-frequency repair and high-frequency reconstruction networks, respectively. We performed the following ablation experiments to verify their efficiency. As illustrated in Table 5, a total of four sets of experiments were performed: (i) default model (proposed model), (ii) removal of SA alone from low-frequency repair network, (iii) removal of FEB alone from high-frequency reconstruction network, and (iv) removal of both SA and FEB. The results show that the evaluation metrics using the default model are the best, and consider removing a module alone decreases the effect of each module accordingly. Furthermore, the effect was the worst in the case of removing both SA and FEB. It was proven that the proposed SA and FEB were indispensable in the proposed network.

TABLE 4. Quantitative results for various evaluation metrics at different wavelet decomposition scales, bold font represents the best performance.

Wavelet scale	PSNR ↑	SSIM ↑	LPIPS ↓	DeltaE ↓	RT ↓
l=1	29.4668	0.9243	0.0328	3.8933	**0.0249s**
l=2	29.8234	0.9364	0.0282	3.3725	0.0481s
l=3	**30.0850**	0.9437	**0.0252**	**2.8494**	0.0562s
l=4	29.9246	**0.9449**	0.0285	3.0475	0.0680s
l=5	29.6816	0.9413	0.0340	3.3182	0.0884s

TABLE 5. In a blation studies of the effectiveness of the simplified attention (SA) and feature extraction block modules (FEB), bold font represents the best performance.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	DeltaE ↓
Default	**30.0850**	**0.9437**	**0.0252**	**2.8494**
w/o SA	29.6868	0.9275	0.0641	3.8542
w/o FEB	29.7714	0.9329	0.0271	3.1169
w/o SA, FEB	27.9486	0.8843	0.0709	6.3844

V. CONCLUSION

This paper put forward a deep-learning model for LLIE. Our key idea is to construct a frequency pyramid of low-light images in terms of a wavelet transform, starting from the global color information of the image and gradually enhancing the image details. To this end, for different frequency components, we constructed a low-frequency restoration network (LRN) and a high-frequency reconstruction network (HRN) to enhance the low-frequency and high-frequency information, respectively. Moreover, we introduce a simplified attention mechanism (SA) and a feature extraction block (FEB) in different networks and demonstrate their effectiveness in ablation experiments. Additionally, we constructed a new dataset consisting of 4500 multi-scale expo-sure image pairs to compensate for the shortcomings of the existing datasets. This enables the proposed model to recover precise details, sharp contrasts, and vivid colors in complex low-light scenes. Extensive experiments prove that our proposed method produces more convincing results than the existing LLIE approaches. In the future work, we will address the problem of unprocessed sensor noise and challenging imaging under extremely dark conditions.

REFERENCES

[1] N. Sharma, V. Kumar, and S. K. Singla, "Single image defogging using deep learning techniques: Past, present and future," *Arch. Comput. Methods Eng.*, vol. 28, no. 7, pp. 4449–4469, Dec. 2021.

[2] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and J.-B. Huang, "Single-image HDR reconstruction by learning to reverse the camera pipeline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 1648–1657.

[3] K. Singh, R. Kapoor, and S. K. Sinha, "Enhancement of low exposure images via recursive histogram equalization algorithms," *Optik*, vol. 126, no. 20, pp. 2619–2625, Oct. 2015.

[4] P. Gupta, J. Kumare, U. Singh, and R. Singh, "Histogram based image enhancement techniques: A survey," *Int. J. Comput. Sci. Eng.*, vol. 5, no. 6, pp. 475–484, 2017.

[5] P. Babakhani and P. Zarei, "Automatic gamma correction based on average of brightness," *Adv. Comput. Sci., Int. J.*, vol. 4, pp. 156–159, Nov. 2015.

[6] S. Rahman, M. M. Rahman, M. Abdullah-Al-Wadud, G. D. Al-Quaderi, and M. Shoyaib, "An adaptive gamma correction for image enhancement," *EURASIP J. Image Video Process.*, vol. 2016, no. 1, pp. 1–13, Dec. 2016.

[7] E. H. Land, "An alternative technique for the computation of the designator in the retinex theory of color vision," *Proc. Nat. Acad. Sci. USA*, vol. 83, no. 10, pp. 3078–3080, 1986.

[8] L. Wang, G. Fu, Z. Jiang, G. Ju, and A. Men, "Low-light image enhancement with attention and multi-level feature fusion," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Shanghai, China, Jul. 2019, pp. 276–281.

[9] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comput.*, vol. 19, no. 90, p. 297, Apr. 1965.

[10] J. Huang, Y. Liu, F. Zhao, K. Yan, J. Zhang, Y. Huang, M. Zhou, and Z. Xiong, "Deep Fourier-based exposure correction network with spatial-frequency interaction," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 163–180.

[11] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Salt Lake City, UT, USA, Jun. 2018, pp. 886–88609.

[12] S. M. Pizer, R. E. Johnston, J. P. Ericksen, B. C. Yankaskas, and K. E. M´uller, "Contrast-limited adaptive histogram equalization: Speed and effectiveness," in *Proc. 1st Conf. Visualizat. Biomed. Comput.*, Atlanta, GA, USA, Sep. 1990, pp. 337–345.

[13] M. Kaur, J. Kaur, and J. Kaur, "Survey of contrast enhancement techniques based on histogram equalization," *Int. J. Adv. Comput. Sci. Appl.*, vol. 2, no. 7, pp. 137–141, 2011.

[14] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based hdr reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–11, Nov. 2012, doi: 10.1145/2366145.2366222.

[15] H. Zhang, E. Zhu, and Y. Wu, "High dynamic range image generating algorithm based on detail layer separation of a single exposure image," *Acta Automatica Sinica*, vol. 45, no. 11, pp. 2159–2170, 2019.

[16] R. Fattal, D. Lischinski, and M. Werman, "Gradient domain high dynamic range compression," in *Proc. 29th Annu. Conf. Comput. Graph. Interact. Techn.*, Jul. 2002, pp. 249–256, doi: 10.1145/566570.566573.

[17] I. Merianos and N. Mitianoudis, "A hybrid multiple exposure image fusion approach for HDR image synthesis," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Chania, Greece, Oct. 2016, pp. 222–226.

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

[19] X. Guo, Y. Li, and H. Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, Feb. 2017.

[20] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An experiment-based review of low-light image enhancement methods," *IEEE Access*, vol. 8, pp. 87884–87917, 2020.

[21] A. Garg, X.-W. Pan, and L.-R. Dung, "LiCENt: Low-light image enhancement using the light channel of HSL," *IEEE Access*, vol. 10, pp. 33547–33560, 2022.

[22] X. Zhang and X. Wang, "MARN: Multi-scale attention retinex network for low-light image enhancement," *IEEE Access*, vol. 9, pp. 50939–50948, 2021.

[23] W. Huang, Y. Zhu, and R. Huang, "Low light image enhancement network with attention mechanism and retinex model," *IEEE Access*, vol. 8, pp. 74306–74314, 2020.

[24] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.

[25] R. Al Sobbahi and J. Tekli, "Low-light image enhancement using image-to-frequency filter learning," in *Proc. Int. Conf. Image Anal. Process.*, 2022, pp. 693–705.

[26] 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 Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[27] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022, doi: 10.1109/TPAMI.2021.3063604.

[28] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Orleans, LA, USA, Jun. 2022, pp. 5627–5636.

[29] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "SNR-aware low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Orleans, LA, USA, Jun. 2022, pp. 17693–17703.

[30] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Paris, France, Oct. 2023, pp. 12504–12513.

[31] L. Chi, G. Tian, Y. Mu, L. Xie, and Q. Tian, "Fast non-local neural networks with spectral residual learning," in *Proc. 27th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2019, pp. 2142–2151.

[32] Y. Rao, W. Zhao, Z. Zhu, J. Lu, and J. Zhou, "Global filter networks for image classification," 2021, *arXiv:2107.00645*.

[33] Q. Xu, R. Zhang, Y. Zhang, Y. Wang, and Q. Tian, "A Fourier-based framework for domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 14378–14387.

[34] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[35] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 41, no. 7, pp. 909–996, Oct. 1988.

[36] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30, no. 1, p. 3.

[37] A. Odena, V. Dumoulin, and C. Olah. *Deconvolution and Checkerboard Artifacts*. Accessed: Oct. 17, 2016. [Online]. Available: https://distill.pub/2016/deconv-checkerboard/?ref=mlq-ai
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.
[39] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1800–1807.
[40] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Cross attention network for few-shot classification," in *Proc. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[42] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D: Nonlinear Phenomena*, vol. 60, nos. 1–4, pp. 259–268, Nov. 1992.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[44] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[45] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *Proc. CVPR*, Colorado Springs, CO, USA, Jun. 2011, pp. 97–104.
[46] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation," in *Proc. 19th IEEE Int. Conf. Image Process.*, Orlando, FL, USA, Sep. 2012, pp. 965–968.
[47] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.
[48] V. Vonikakis, D. Chrysostomou, R. Kouskouridas, and A. Gasteratos, "Improving the robustness in feature detection by local contrast enhancement," in *Proc. IEEE Int. Conf. Imag. Syst. Techn.*, Jul. 2012, pp. 158–163.
[49] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Comput. Vis. Image Understand.*, vol. 178, pp. 30–42, Jan. 2019.
[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Lake City, UT, USA, Jun. 2018, pp. 586–595.
[51] D. H. Brainard, "Color appearance and color difference specification," *Sci. Color*, vol. 2, nos. 191–216, p. 5, 2003.
[52] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[53] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Orleans, LA, USA, Jun. 2022, pp. 5891–5900, doi: 10.1109/CVPR52688.2022.00581.
[54] 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," 2023, *arXiv:2302.11831*.
[55] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding, "YOLOv10: Real-time end-to-end object detection," 2024, *arXiv:2405.14458*.

YANGJUN XIANG received the bachelor's degree in environmental engineering from Henan University of Urban Construction, in 2022. He is currently pursuing the M.S. degree in digital media technology with Hangzhou Dianzi University. His research interests include image enhancement, computational imaging, and deep learning.

GENGSHENG HU received the M.S. degree in signal and information processing from Xi'an University of Technology, in 1997. He is currently a Professor with Hangzhou Dianzi University. His research interests include image processing, color management, and intaglio printing.

MEI CHEN received the B.S. degree in printing enterprise management from Beijing Institute of Graphic Communication, in 1991, and the Ph.D. degree in printing and packaging technology and equipment from Xi'an University of Technology, in 2013. She is currently an Associate Professor with Hangzhou Dianzi University. Her current research interests include print media technology, business management, and intaglio printing.

MAHMOUD EMAM received the B.S. and M.S. degrees in mathematics and in computer science from Menoufia University, Egypt, in 2006 and 2012, respectively, and the Ph.D. degree in computer science and technology from Harbin Institute of Technology, Harbin, China, in 2017. From 2013 to 2017, he was a Research Associate with the Research Institute of Information Countermeasure Techniques (ICT), School of Computer Science and Technology, Harbin Institute of Technology. He is currently an Assistant Professor in computer science with Menoufia University. He is the author or co-author of many publications and a referee of many refereed international journals and conferences. His research interests include the digital forensics, multimedia security, computer vision, machine learning, and pattern recognition. He received two best paper awards from the IEEE and Springer conferences. His awards and honors include the Best Foreigner Ph.D. Student Award and the 2016 Outstanding International Ph.D. Student Excellence Award by Chinese Scholarship Council, China, in 2016.