Computers and Electrical Engineering 106 (2023) 108608

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

# A cross-scale framework for low-light image enhancement using spatial–spectral information☆

Yichi Zhang, Hengyu Liu, Dandan Ding *

*School of Information Science and Technology, Hangzhou Normal University, Hangzhou, 311121, China*

ARTICLE INFO

*Keywords:*
Low-light
Transformer
Cross-scale
Fast Fourier Convolution
Image enhancement

ABSTRACT

This paper presents a spatial–spectral low-light enhancement approach based on the neural network. Considering that the Image Signal Processor (ISP) which non-linearly maps RAW data to RGB images may introduce additional noise and artifacts, we propose to conduct the enhancement task directly on RAW data. To this end, we propose a cross-scale framework to map the input low-light RAW data to visually pleasing RGB images. The cross-scale framework consists of three branches for low-level, mid-level, and high-level representations of input images. We embed paired Fast Fourier Convolution (FFC) and Transformer in each level for global and local feature analysis and aggregation. Specifically, the FFC enlarges the receptive field of our neural network, exploring the pixel correlations in the spectral space; the following Transformer uses self-attention for feature selection and aggregation in the spatial space. As a result, spatial–spectral information within an image is jointly utilized for the final fusion. Experimental results demonstrate that the proposed approach significantly outperforms state-of-the-art low-light enhancement methods in both full reference assessment metrics, including PSNR, MPSNR, and SSIM, and no-reference metrics, such as NIMA. Moreover, the proposed method produces more aesthetically pleasing RGB images than other methods.

## 1. Introduction

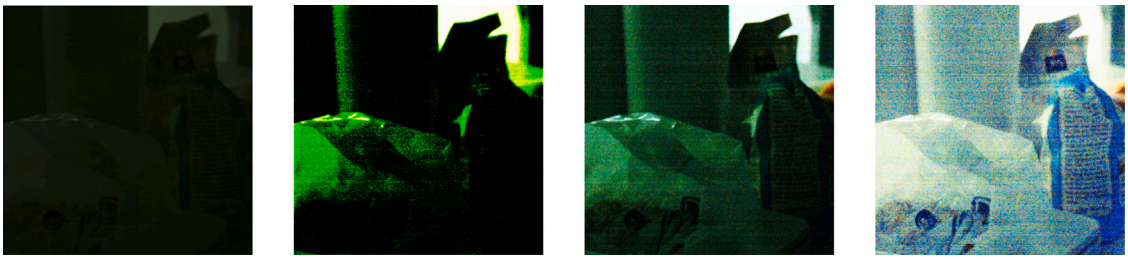
Imaging in the low-light environment is important for applications including security, autonomous driving, robotics. However, images captured in low-light environments often suffer from poor aesthetic quality because of various degradation such as noise, low visibility, and color cast which not only compromises aesthetic quality but also causes inaccurate object/face recognition. To solve this issue, image enhancement solutions are required for images taken under low-light conditions to recover a visually pleasing image without the color cast, noise, etc., for human viewing.

Image enhancement has been widely used in various practical applications, such as medical diagnosis, underwater exploration, and highway detection. The most commonly used method is RGB–RGB mapping: given a low-light RGB image, a specific algorithm is applied to generate its enhanced RGB counterpart. Typical methods include the adaptive Histogram Equalization [1] that stretches the dynamic range of an image globally or locally and the Retinex model [2] that decomposes images into reflectance and illumination for respective enhancement. These methods improve the image quality objectively or subjectively under some prior knowledge or assumptions but may fail in scenarios where these prior knowledge or assumptions do not hold true. By contrast, learning-based methods [3–8] adopt a data-driven method for end-to-end learning. Compared with those handcrafted traditional algorithms, learning-based methods have successfully achieved superior performance.

☆ This paper is for special section VSI-eltec. Reviews were processed by Guest Editor Dr. Xiaohang Wang and recommended for publication.

* Corresponding author.
*E-mail address:* DandanDing@hznu.edu.cn (D. Ding).

https://doi.org/10.1016/j.compeleceng.2023.108608
Received 30 August 2022; Received in revised form 14 January 2023; Accepted 20 January 2023
Available online 27 January 2023



(a) RAW image (b) AWNet [17] (c) PyNET [15] (d) PyNET-CA [16]

**Fig. 1.** Examples of directly applying conventional neural ISPs to low-light RAW inputs to generate RGB images.

Another category of image enhancement methods [9–13] start directly from the low-light RAW data and generate corresponding high-quality RGB images, so-called as RAW-RGB. The RAW data are directly from the image sensor and contain light intensity values captured from a scene. As can be seen, the RAW image contains more native data compared to the RGB image, which is able to provide more useful information for the enhancement task. On the other side, the RAW data are also easier to be represented because they are linear to exposure levels [14]. Relatively, for the RGB–RGB manner, we need to map the low-light RAW data to low-light RGB image by applying the Image Signal Processor (ISP), which is a complex non-linear mapping process including demosaicing, white balance, tone mapping, etc. This not only loses certain significant details but also amplifies the noise and useless information from sensors, making the RGB image more difficult to be enhanced. Clearly, RAW-RGB is more beneficial for modeling, analysis, and enhancement. In this paper, we focus on the learning-based RAW-RGB mapping for low-light image enhancement.

Essentially, RAW-RGB involves not only enhancement but also ISP function. In the past few years, numerous neural ISPs were developed to convert the RAW data into RGB images, such as PyNET [15], PyNET-CA [16], and AWNet [17]. Most of these works are devised for images captured under normal illumination in the natural environment and could be used to replace traditional IPSs. Since the low-light scenarios are not considered in their designs, they usually fail in low-light conditions and suffer from serious color cast and distortion. As illustrated in Fig. 1 where a typical low-light image is processed by prevalent neural ISPs. All the RGB images generated exhibit excessive noise and chromatic aberration, e.g., after enhancement by AWNet [17], the shape of the object can barely be seen. This reveals that conventional neural IPSs are limited to handling low-light RAW images.

To this end, we propose a dedicated neural network-based approach to improve the quality of images captured in a low-light environment. We first build a cross-scale framework that consists of three branches for low-level (edges, blobs, etc.), mid-level (meaningful shapes), and high-level (semantics) feature extraction. Then, features from different scales are collected and fused for final enhancement. In each scale, we embed paired Fast Fourier Convolution (FFC) [18] and Transformer for spatial–spectral information abstraction and aggregation. Specifically, the FFC is first applied to transform the input feature map to the spectral domain where each element represents a weighting sum of all features in the input feature map; after applying convolutions in the spectral domain for feature capturing, we transform features back to the spatial domain. Immediately, a Transformer is followed to generate attention maps for captured features. The cross-scale framework together with the paired FFC and Transformer effectively enlarges the receptive field and utilizes global and local features jointly for the enhancement purpose. As an extension of our previous work [19], this work provides a detailed network description and conducts in-depth studies on the contribution of each modular component and each item in the loss function; exhausted experiments are also applied to reveal the effectiveness of paired FFC and Transformer. Moreover, we devise an advanced version with higher performance. Extensive experiments confirm the effectiveness of the proposed cross-scale framework using paired FFC and Transformer.

The highlights of this work are summarized below.

- We propose a learning-based cross-scale framework to directly map the RAW data captured in the low-light environment to subjectively pleasing RGB images. Within the proposed framework, features of different resolutions are captured for low-level, mid-level, and high-level representation and combined for attention-aware enhancement.
- We utilize paired FFC and Transformer to extract latent features in each scale of our framework. By transforming features into the spectral domain, pixel correlations are effectively exploited in a large receptive field; Transformer further deploys the self-attention mechanism to aggregate and strengthen useful features.
- The cross-scale framework, together with the paired FFC and Transformer collects both global and local information for final enhancement. Experimental results demonstrate that the proposed method remarkably outperforms state-of-the-art low-light enhancement works in both full-reference assessment metrics such as PSNR, MPSNR, and SSIM, and no-reference metrics, such as NIMA. Moreover, our method attains much higher perceptual quality than other methods.

The remainder of this paper is organized as follows. Section 2 introduces related works. The proposed method is detailed in Section 3. Section 4 shows our experiment settings and results. Finally, Section 5 concludes this work.

## 2. Related work

This section reviews relevant studies on low-light image enhancement and Transformers.

### 2.1. Learning-based low-light image enhancement

As summarized in [20], most state-of-the-art learning-based methods enhance low-light images following the RGB–RGB manner. Ren et al. [21] proposed a hybrid network consisting of a content stream based on an encoder–decoder network and an edge stream based on a spatially variant recurrent neural network (RNN) to simultaneously learn the global content and salient structure of an image in a unified network. Besides, some works attempted to combine the Retinex theory to design deep neural networks. Chen et al. [3] proposed to enhance illumination and reflectance by using separate sub-networks. Zhang et al. [6] devised three sub-networks for layer decomposition, reflectance recovery, and illumination adjustment. Fan et al. [22] combined the semantic segmentation and Retinex models to further improve enhancement performance in real-world situations, making full use of prior information on image semantics to guide the enhancement of illumination and reflection components.

In addition to the supervised learning above, unsupervised learning and Zero-Shot Learning are also developed. To tackle the issue that supervised learning-based models may overfit, an unsupervised learning method called EnligthenGAN [4] was proposed by regularizing the unpaired training using the information extracted from the input itself. Zero-DCE [7] treated the low-light enhancement as a curve estimation task, taking low-light images as input and generating high-order curves as its output. These curves were used to make pixel-level adjustments to the dynamic range of the input to obtain an enhanced image. It was updated to a faster and more lightweight version, called Zero-DCE++ [8] later.

The methods aforementioned are effective in improving low-light enhancement, but their performance is restricted due to the inherent limitations of the RGB format. For example, low-light images tend to contain a large amount of noise, which becomes more complex and difficult to handle after conventional ISP processing. Liu et al. [23] proposed a flexible and practical way by mapping the RGB image back to RAW image and enhancing the mapped RAW image. They trained a RAW-guiding Exposure Enhancement Network (REENet) to accomplish the RGB to RAW mapping and only RGB images are needed in the test. This is a typical RGB-RAW-RGB method. However, the RAW image from the REENet is lossy (compared with the ground truth RAW data) and introduces additional noise which will be amplified in the following RGB to RGB procedure, further influencing the performance. On the other hand, since there are abundant RAW images currently, we can skip the RGB to RAW step and apply the RAW-RGB mapping directly.

Numerous works have attempted to leverage the RAW data for low-level vision tasks. For example, CycleISP [24] modeled the camera imaging pipeline in the forward and reverse directions, generating an arbitrary number of realistic image pairs in RAW and sRGB space for denoising. Later, the RAW-RGB method is also adopted for low-light enhancement. Many methods are proposed for designing a low-light ISP network starting directly from RAW data. Chen et al. [9] trained a fully convolutional network (FCN) to perform the entire ISP pipeline, which operated not on the normal sRGB images produced by a traditional ISP, but on the raw sensor data. Paras et al. [12] proposed a deep neural network based on residual learning for end-to-end extreme low-light image denoising, providing better color reproduction and preserving more detailed texture information. Eli et al. [25] proposed DeepISP, an end-to-end deep neural model of the complete camera ISP pipeline that learned the mapping from the original low-light mosaic image to the final visually convincing image. Zhu et al. [10] proposed a two-stage approach, called edge-enhanced multi-exposure fusion network (EEMEFN): first, a multi-exposure fusion module was used to address high contrast and color bias, and secondly, an edge enhancement module was introduced to refine the initial image with the help of edge information. Zhang et al. [11] proposed an attention-based neural network by employing attention strategies (i.e., spatial attention and channel attention) to suppress unwanted chromatic aberrations and noise. In addition, a pooling layer, called the inverted shuffle layer, was used to select useful information from the previous features. Ahmet Serdar et al. [13] exploited continuous shooting to improve performance by proposing a coarse-to-fine network structure, progressively producing high-quality output. The coarse network predicted the low-resolution, denoised RAW image and then fed it to the fine network to recover fine-scale details and realistic textures. Ignatov et al. [15] proposed PyNET, a novel pyramidal CNN architecture designed for fine-grained image restoration. Dai et al. [17] introduced the AWNet that incorporated wavelet transform to recover favorable image details from RAW information, which achieved a larger receptive domain while remaining efficient computational cost.

### 2.2. Transformer in vision tasks

ViT [26] first introduced the Transformer in vision tasks and achieved impressive performance. ViT divided images into tokens that are then fed into the Transformer model for exploiting long-range dependencies using global attention. Since then, the Transformer-based models have been widely applied in various computer vision tasks including image/video classification, object detection, semantic segmentation, etc. Apart from these high-level tasks, in the field of low-level vision, transformers have also made great progress. Lu et al. [27] proposed an Efficient Super-Resolution Transformer (ESRT) for using image super-resolution. It consists of a Lightweight Convolutional Neural Network (CNN) Backbone (LCB) and a Lightweight Transformer Backbone (LTB). Zamir et al. [28] proposed a Transformer model by modifying multi-head self-attention and feed-forward networks, which can capture long-range pixel interactions while still being applicable to large images, achieving state-of-the-art results on several image restoration tasks. Wang et al. [29] constructed a hierarchical encoder–decoder network for image restoration using a Locally enhanced Window (LeWin) Transformer.

The studies above demonstrate the impressive performance of Transformers for different vision tasks. Inspired by these works, we devise a cross-scale network and combine Fast Fourier Convolution with LeWin Transformer, utilizing both global features and local self-attention to fully explore the spatial–spectral correlations.

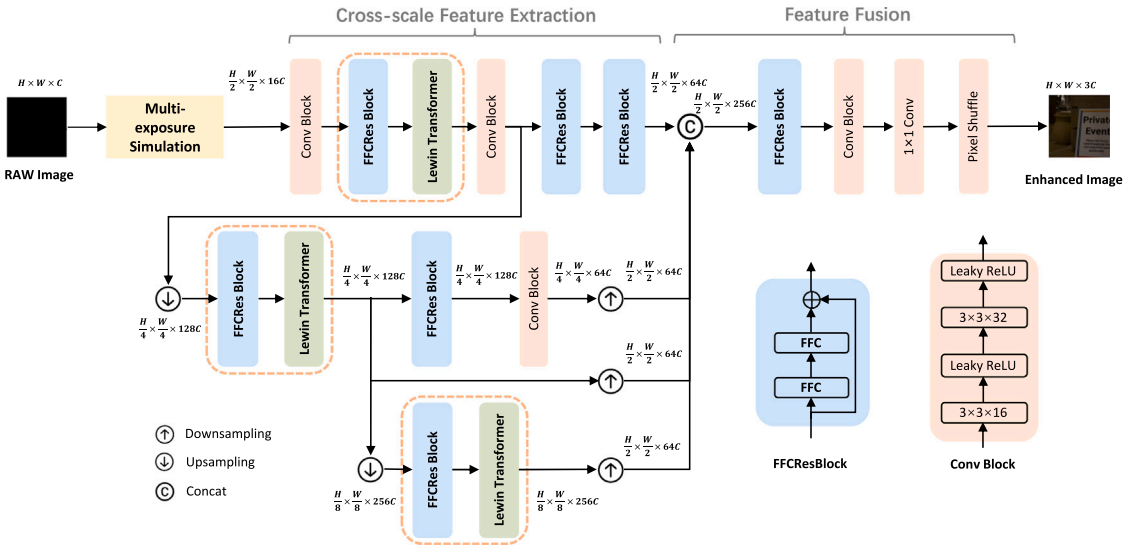



**Fig. 2.** The proposed cross-scale framework which consists of three branches for low-level, mid-level, and high-level feature extraction. In each scale, paired FFCRes Block (each FFCRes Block contains two FFCs) and Transformer is applied for spectral–spatial feature aggregation. At last, features of the three scales are concatenated for fusion.

## 3. Proposed low-light enhancement framework

### 3.1. Proposed cross-scale architecture

As illustrated in Fig. 2, the proposed low-light enhancement framework consists of three core components: the Multi-exposure Simulation module, the Cross-scale Feature Extraction module, and the Feature Fusion module. Given a low-light RAW image, the Multi-exposure Simulation module first applies different exposure levels to the RAW image to generate a set of enhancement candidates. Then, all these candidates are sent to the Cross-scale Feature Extraction module which has three branches to explore both spatial and spectral correlations by downsampling the feature maps to different resolutions. At last, multi-level features are aggregated and fused through the Feature Fusion module for final enhancement. Details of each module will be described below.

### 3.2. Multi-exposure Simulation

We incorporate a linear Multi-exposure Simulation module before the cross-scale network. Since the RAW image is linear to exposure levels [14], it is straightforward to simply multiply the RAW data by different exposure factors $\lambda$ to mimic different exposure levels. This will result in a set of enhancement candidates and allow the subsequent network to capture rich information from these candidates.

Specifically, given a RAW image $\mathbf{R} \in \mathbb{R}^{C \times H \times W}$, we first linearly interpolate $\mathbf{R}$ to produce a color image $\widetilde{\mathbf{I}} \in \mathbb{R}^{4C \times \frac{H}{2} \times \frac{W}{2}}$ using the demosaicing operation. Afterwards, we multiply $\widetilde{\mathbf{I}}$ by exposure factors $\lambda_i$, $i = 1, 2, \ldots, k$, generating $\mathbf{I}_i$ as candidates for network processing, described as follows:

$$\mathbf{I}_i = \lambda_i \times F_d(\mathbf{R}), \tag{1}$$

where $F_d(\cdot)$ denotes the demosaicing operation. In this work, $k$ is set as 4 and weighting parameters $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4$ are set to 0.8, 1.0, 1.2, and 1.5, respectively.

### 3.3. Cross-scale feature extraction

As described in Fig. 2, the entire *Cross-scale framework* is comprised of three branches by gradually downsampling the feature maps to three resolutions. These branches are used for respective low-level (edges, blobs, etc.), mid-level (meaningful shapes), and high-level feature extraction. We first feed the cascaded $\mathbf{I}_i$ into the low-level branch. Here a Convolutional Block (Conv Block) is conducted, followed by a pair of FFC Residual Block (FFCRes Block) and a Transformer. Then, another Conv Block is applied again to generate feature maps with the size of $H/2 \times W/2 \times 64$. These feature maps on the one hand continue to be processed by two stacked FFCRes Blocks in the low-level branch and on the other hand, are downsampled to the size of $H/4 \times W/4 \times 128$ for mid-level processing. Similarly, after a pair of FFCRes Block and Transformer, the mid-level feature maps are downsampled again to the size of $H/8 \times W/8 \times 256$ for the high-level processing. Meanwhile, the mid-level features still go across the FFCRes Block

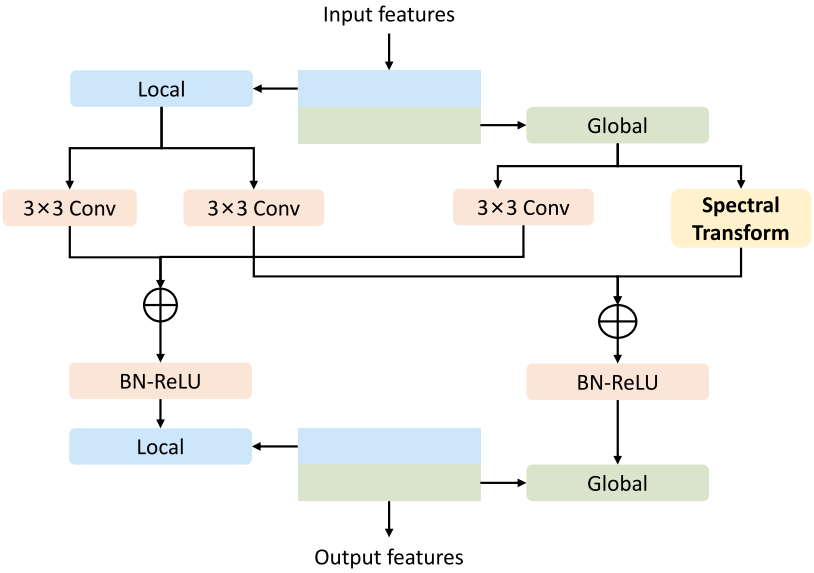



**Fig. 3.** Structure of FFC which is divided into two groups: one for local and the other for global feature extraction.

for feature fusion. At last, these mid-level and high-level feature maps are uniformly upsampled to the size of $H/2 \times W/2 \times 64$ for concatenation, forming a hybrid $H/2 \times W/2 \times 256$ features for the final Feature Fusion.

Specifically, the structure of the Conv Block is shown in Fig. 2. It consists of two convolutional layers and each layer follows a LeakyReLU for activation.

As seen in Fig. 2, a pair of *FFCRes Block* and *Transformer* is embedded in each scale and contribute significantly to the enhancement performance. The FFCRes Block transforms spatial features into the spectral domain for convolution. As such, a value in the spectral domain is a weighted representation of all spatial data, which implicitly expands the network receptive field. Furthermore, after FFC follows a Transformer block to collect and strengthen the important features. As a result, spatial–spectral information is jointly utilized. Their respective designs are presented below.

**Fast Fourier Convolution Residual Block.** As shown in Fig. 2, two *FFC* modules constitute an FFCRes Block, which helps two distant locations to connect effectively and efficiently. Within each FFC (see Fig. 3), the input paramilitary features are divided into two groups: one for local and the other for global feature extraction. Specifically, for the global group, input features are transformed into the Fourier domain for further enhancement through the *Spectral Transform (ST)* block which implicitly covers a large receptive field; for the local branch, two parallel $3 \times 3$ convolutional layers are deployed, focusing feature perception within the local field. Afterwards, the global and local information are mixed together for subsequent processing.

The *Spectral Transform* in FFC, as described in Fig. 4, still follows the local–global principle, where the global group (the left branch in Fig. 4) operates on the entire image, and the local group (the right branch in Fig. 4) splits an image into four smaller feature maps for processing. Specifically, a $1 \times 1$ convolution is first applied to reduce the number of channels for computational complexity saving. For the global branch, a 2-D Fast Fourier Transform (FFT) is used to convert the whole feature map to the spectral domain where a $1 \times 1$ convolution is conducted for global updating, followed by batch normalization and ReLU operations; after that, features are converted back to the spatial space. For the local branch, we focus more on the local information by slicing every input feature map into four pieces. For each piece, the 2D FFT, $1 \times 1$ convolution, and inverse 2D FFT are conducted. Finally, we repeat the output feature maps four times to produce an output with the same size as the input feature map.

**LeWin Transformer.** The features from FFCRes Block contribute differently to the final output. To this end, we need to emphasize the important features and attenuate the useless ones. Then, how to evaluate the contribution of each feature is a problem. In this paper, the self-attention mechanism in Transformer is introduced to fulfill this purpose. However, there are some challenges when applying conventional Transformers [26] for image restoration: (1) The global computation of self-attention imposes huge computational costs; (2) Conventional transformers have limitations in capturing local dependencies while the local context information is essential for image restoration tasks. Inspired by recent progress on low-level vision Transformers [28], we introduce LeWin Transformer [29], which can effectively address these issues through two core designs: Window-based Multi-head Self-Attention (W-MSA) and Locally-enhanced Feed-Forward Network (LeFF) (see Fig. 5). On the one hand, FFCResBlock works in a large receptive field and captures long-range dependencies spectrally. On the other hand, W-MSA and LeFF perform on local windows and capture local context correlations spatially. As such, local–global features are aggregated for a better enhancement effect.

*Window-based Multi-head Self-Attention (W-MSA):* Given feature maps $X \in \mathbb{R}^{H \times W \times C}$, where $H$ and $W$ are the height and width of the map, respectively, and $C$ is the channel number, we firstly partition $X$ into non-overlapping windows of fixed size $M \times M$.

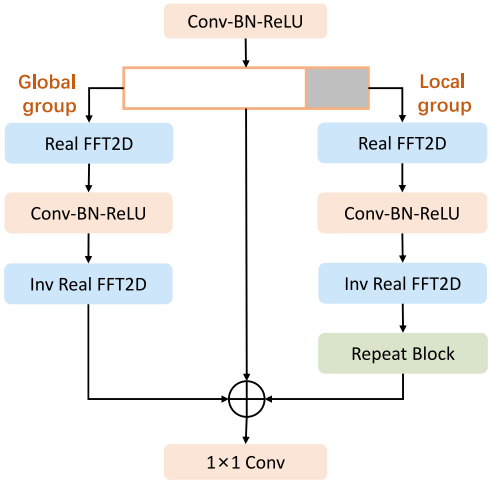



**Fig. 4.** Spectral Transform (ST) used in the FFC to transform features into spectral space for local and global feature perception.

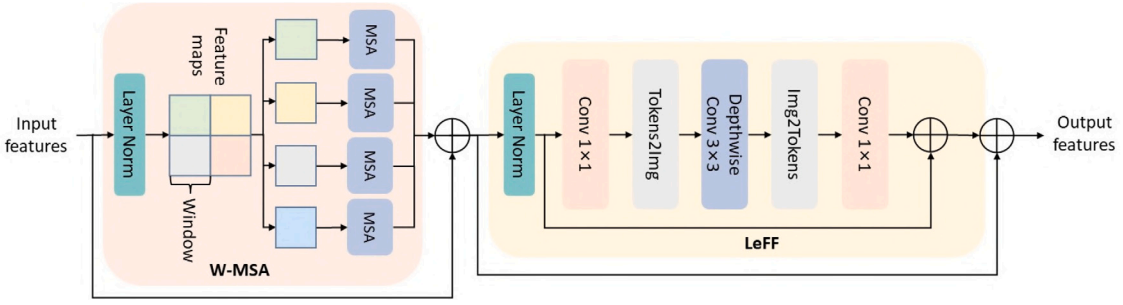



**Fig. 5.** The structure of LeWin Transformer used in this work. It adopts the Multi-head Self-Attention (W-MSA) and Locally-enhanced Feed-Forward Network (LeFF) on a local fixed-size window.

Next, self-attention is applied to features inside each window. Suppose the number of heads is $k$ and the head dimension is $d_k = C/k$. Then the self-attention of the $k$th head in the non-overlapping window is calculated as:

$$Y_k^i = SoftMax(\frac{X^i W_k^Q (X^i W_k^K)^T}{\sqrt{d_k}}) X^i W_k^V \tag{2}$$

where $X_i$ is the feature from window $i$. $W_k^Q$, $W_k^K$, $W_k^V$ represent the weights of the queries, keys, and values layers for the $k$th head, respectively. The head outputs are then concatenated to pass through a linear layer to produce the final output.

*Locally-enhanced Feed-Forward Network (LeFF):* The feed-forward network in standard Transformers has limited ability to exploit the local context. In fact, adjacent pixels are important references for image restoration. To overcome this problem, LeFF [29] is proposed. LeFF adds a depthwise convolution block to the feed-forward network in the Transformer structure. First, a linear projection layer is applied to each token to increase its feature dimensionality. Next, tokens are reshaped into a 2D feature map and a $3 \times 3$ depthwise convolution is used to capture local information. Then, the features derived are panned back to tokens and the number of channels is reduced by another linear layer to match the dimensions of the input.

### 3.4. Feature Fusion

In the Feature Fusion stage, after collecting features from multiple scales, we obtain $W/2 \times H/2 \times 256$ features, which sequentially go across an FFCResBlock, a Conv Block, and a $1 \times 1$ convolutional layer. Finally, we shuffle the produced pixels and obtain the enhanced RGB image.

### 3.5. Loss function

When training the proposed network, we combine four loss items to form a hybrid loss, including the conventional Mean Absolute Error (MAE) loss $\mathcal{L}_{MAE}$, the Structural Similarity Index Measure (SSIM) loss $\mathcal{L}_{SSIM}$, the Gradient Difference Loss (GDL) $\mathcal{L}_{GDL}$, and the FFT loss $\mathcal{L}_{FFT}$.

Since we leverage the FFC in our architecture, spectral loss is important to supervise the learning. We hence introduce the FFT loss to preserve texture details as

$$\mathcal{L}_{FFT}(E,R) = \frac{1}{N}\sum_{i=1}^{n} \|F_E - F_R\|, \tag{3}$$

where $F_E$ represents the enhanced image after FFT and $F_R$ represents the reference image after FFT.

Combining all loss metrics, our final loss function can be expressed as:

$$\mathcal{L} = a \times \mathcal{L}_{MAE} + b \times \mathcal{L}_{SSIM} + c \times \mathcal{L}_{GDL} + d \times \mathcal{L}_{FFT}, \tag{4}$$

where $a$, $b$, $c$, and $d$ are linear combination parameters. Inspired by the SID implementation [9] where the use of MAE loss achieves a significant enhancement effect, we set parameter $a = 0.6$ here to emphasize the MAE contribution. The weights $b$ and $c$ are both set to 0.2 and $d$ is set to 0.01.

## 4. Experiments and discussions

This section presents the quantitative and qualitative performance of the proposed method. Ablation studies are conducted to provide more evidence on the effectiveness of the proposed method.

### 4.1. Implementation details

**Dataset.** We use the SID dataset [9] to evaluate the performance of our method. The SID dataset [9] contains 5094 short-exposure and 424 long-exposure RAW images taken by Sony and Fuji cameras in extremely low-light conditions. Each scenario has a low-light RAW image with different exposure times and a long exposure RAW image as the ground truth, where the short exposure time includes 0.1 s, 0.04 s, and 0.03 s, and the corresponding long exposure time is 10 s, 30 s The image resolution is 4240 × 2832 for the Sony set and 6000 × 4000 for the Fuji set.

Moreover, an RGB dataset called LOL [3] is also used for performance evaluation. The LOL dataset [3] contains 500 pairs of low-light/normal-light RGB images with the size of 600 × 400, of which 485 pairs are for training and 15 pairs are for testing.

**Training and Testing.** The training and testing processes are both conducted on a computer with an Intel i7 9770k CPU at 3.6 GHz, 64 GB of main memory, and a 3090 GPU. In the training, for the SID dataset, we randomly crop 512 × 512 patches from low-light RAW images as input. Meanwhile, the corresponding long-exposure RAW images are cropped and processed by LibRaw to produce the RGB image patches as ground truth. We train the network with 4000 epochs on the SID dataset. During training, we use the Adam optimizer and set the initial learning rate to 1e−4. The learning rate drops to 2e−5 after 2000 epochs and to 1e−5 after 3000 epochs. The whole training process takes about 100 h.

For the LOL dataset, we randomly crop 256 × 256 low-light RGB patches as input and use their normal RGB counterparts as ground truth. We train the network with 1000 epochs. The initial learning rate is also set to 1e−4. After 200 epochs, the learning rate dropped to 2e−5.

In the testing, for the SID dataset, we crop patches of size 1024 × 1024 from the center of short-exposure RAW images as input. For the LOL dataset, we crop 384 × 384 size patches at the center of low-light RGB images as input. Corresponding patches in normal images are served as ground truth.

**Evaluation Metrics.** We employ the traditional metrics, including Peak Signal-to-Noise Ratio (PSNR), Mean Peak Signal-to-Noise Ratio (MPSNR), and the SSIM, as objective evaluation metrics. Also, we adopt the Learned Perceptual Image Patch Similarity (LPIPS), which is more consistent with human perception, to evaluate the perceptual difference between the enhanced and the ground truth images. In addition, the non-reference image assessment metric, Neural Image Assessment (NIMA), which assesses image quality from a subjective point of view, is also used for evaluation. NIMA is able to predict the distribution of human evaluations of images from both direct perception (technical perspective) and attractiveness (aesthetic perspective), with the advantage of being very similar to subjective human scoring. Note that the NIMA measurement first predicts ten scores and then computes the final NIMA as $\mu \pm (\sigma)$, where $\mu$ and $\sigma$ represent the mean and standard deviation of these ten scores, respectively.

### 4.2. Comparison with state-of-the-art methods

We first compare the proposed method with state-of-the-art RAW enhancement methods, including SID [9], Residual [12], EEMEFN [10], ALEN [11], DBP [13], PyNET [15], and Uformer [29] and RGB enhancement methods including Retinex [3], KinD [6], Zero-DCE [7], Zero-DCE++ [8], EnlightenGAN [4], RUAS [5], PyNET [15], and Uformer [29]. Moreover, we implement two versions of the proposed method: one lightweight version using only a pair of FFCRes Block and Transformer in each scale (shown in Fig. 2), termed ours, and the other pairing each FFCRes Block with a Transformer, called ours++.

For all methods, we either directly run their open source code and pre-trained models or strictly implement them following their papers.

**Table 1**
Performance comparison of different methods on Sony and Fuji subsets of the SID dataset (bold is the best and red is the second best).

| Method | SID (Sony) | | | | | SID (Fuji) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | MPSNR ↑ | SSIM ↑ | LPIPS ↓ | NIMA ↑ | PSNR ↑ | MPSNR ↑ | SSIM ↑ | LPIPS ↓ | NIMA ↑ |
| SID [9] | 28.5699 | 29.1699 | 0.7735 | 0.4662 | 4.001 ± (1.741) | 26.5749 | 27.2719 | 0.7041 | 0.4394 | 3.220 ± (1.947) |
| Residual [12] | 27.8762 | 28.6096 | 0.7713 | 0.4697 | 4.007 ± (1.742) | 25.4394 | 26.2429 | 0.6896 | 0.4745 | 3.329 ± (1.925) |
| EEMEFN [10] | 29.1520 | 29.8138 | 0.7817 | 0.4481 | 4.090 ± (1.752) | 27.4357 | 28.1697 | 0.7190 | 0.4194 | 3.525 ± (2.012) |
| ALEN [11] | 29.3872 | 29.8441 | 0.7794 | 0.4701 | 4.052 ± (1.750) | 27.6540 | 28.2464 | 0.7110 | 0.4466 | 3.327 ± (1.977) |
| DBP [13] | 29.5036 | 30.0330 | 0.7736 | **0.2996** | 3.865 ± (1.708) | 28.2053 | 28.7784 | 0.7140 | **0.3510** | 3.496 ± (1.936) |
| PyNET [15] | 29.9632 | 30.4325 | 0.7871 | 0.4443 | 4.071 ± (1.746) | 26.4976 | 27.0653 | 0.6964 | 0.4724 | 3.215 ± (1.986) |
| Uformer [29] | 29.6569 | 30.1641 | 0.7847 | 0.4399 | 4.089 ± (1.751) | 27.8591 | 28.4908 | 0.7190 | 0.4182 | 3.492 ± (2.006) |
| Ours | 30.1851 | 30.6853 | 0.7837 | 0.4424 | **4.092 ± (1.751)** | 28.4339 | 29.0682 | 0.7147 | 0.3627 | **3.729 ± (1.967)** |
| Ours++ | **30.3898** | **30.8630** | **0.7915** | 0.4380 | 4.090 ± (1.751) | **28.8062** | **29.2611** | **0.7215** | 0.4002 | 3.431 ± (1.986) |

↑ implies the higher the better; ↓ implies the lower the better.

**Table 2**
Performance comparison of different methods on the LOL dataset (bold is the best and blue is the second best).

| Method | PSNR↑ | MPSNR↑ | SSIM↑ | LPIPS↓ | NIMA↑ |
|---|---|---|---|---|---|
| Retinex [3] | 17.0583 | 17.2938 | 0.4241 | 0.3641 | 4.853 ± (1.514) |
| KinD [6] | 19.9310 | 20.1367 | 0.8261 | 0.1409 | **5.184 ± (1.575)** |
| Zero-DCE [7] | 14.9733 | 15.1041 | 0.5620 | 0.2799 | 4.791 ± (1.536) |
| Zero-DCE++ [8] | 15.4430 | 15.7684 | 0.5713 | 0.2811 | 4.818 ± (1.543) |
| EnlightenGAN [4] | 17.4749 | 17.6363 | 0.6497 | 0.2611 | 4.806 ± (1.537) |
| RUAS [5] | 16.3196 | 16.4046 | 0.4980 | 0.2060 | 4.887 ± (1.559) |
| PyNET [15] | 21.9441 | 22.1526 | 0.8083 | 0.1507 | 4.917 ± (1.539) |
| Uformer [29] | 21.0046 | 21.2964 | 0.8290 | **0.1155** | 5.052 ± (1.544) |
| Ours | 22.8321 | 23.0818 | 0.8634 | 0.1211 | 5.165 ± (1.562) |
| Ours++ | **22.9230** | **23.2428** | **0.8719** | 0.1314 | 5.148 ± (1.550) |

↑ implies the higher the better; ↓ implies the lower the better.

**Table 3**
Computational complexity comparison across different RAW enhancement methods.

| Method | Parameters (M) | MACs (G) |
|---|---|---|
| SID [9] | 7.76 | 55.04 |
| Residual [12] | 2.57 | 687.57 |
| EEMEFN [10] | 40.71 | – |
| ALEN [11] | 2.68 | 46.62 |
| DBP [13] | 93.97 | 1608.68 |
| PyNET [15] | 47.55 | 1789.89 |
| Uformer [29] | 20.76 | 104.41 |
| Ours | 8.11 | 601.90 |
| Ours++ | 12.71 | 699.54 |

#### *4.2.1. Comparison on the SID and LOL datasets*

Table 1 shows the quantitative comparison with state-of-the-art RAW-RGB methods on the SID dataset, with the best highlighted with bold type and the second best colored in red. As shown, our proposed method achieves the best results in PSNR, MPSNR, SSIM, and NIMA metrics. For example, our PSNR is 0.22 dB higher than the second-best method PyNET on the Sony dataset, MPSNR is 0.25 dB higher, and NIMA also achieves the highest score. Ours++ performs even better: 0.42 dB PSNR, 0.43 dB MPSNR, 0.004 SSIM higher than the PyNET. As for LPIPS, ours++ is the second-best of all, which is next to DBP on the Sony dataset. For the Fuji dataset, our method still performs superior, e.g., having the best results in PSNR, MPSNR, SSIM, and NIMA. The DBP model achieves the best LPIPS on the SID dataset but its visual quality suffers from obvious blurring artifacts, as described in Section 4.3.

Similarly, we compare the proposed method on the RGB dataset LOL, as reported in Table 2. Here, our model is retrained using the RGB dataset. It is observed that the proposed method performs the best in terms of PSNR, MPSNR, and SSIM metrics. Specifically, ours++ is slightly better than ours, and it has 0.97 dB PSNR gains, 1.09 dB MPSNR, and 0.064 SSIM over the second-best PyNET. As for the LPIPS metric, our method is second to Uformer but better than others. For the NIMA, our method is only slightly inferior to KinD. These results confirm the superiority of our proposed method.

#### *4.2.2. Computational complexity comparison*

Moreover, we compare the computational complexity of different models in Table 3. For the model parameter, our and our++ models use 8.1M and 12.71M parameters, respectively, which is much smaller than EEMEFN, DBP, PyNET, and Uformer. Notice that the EEMEFN costs 40.71M, almost 5× parameters than ours, and the DBP model costs 93.97M, almost 12× of ours. Relative to the Residual model, although our model uses more parameters, our GMACs are smaller, i.e., 602 GMACs vs 688 GMACs. We notice that the SID and the ALEN models have much fewer parameters and GMACs than ours, but their enhancement performance

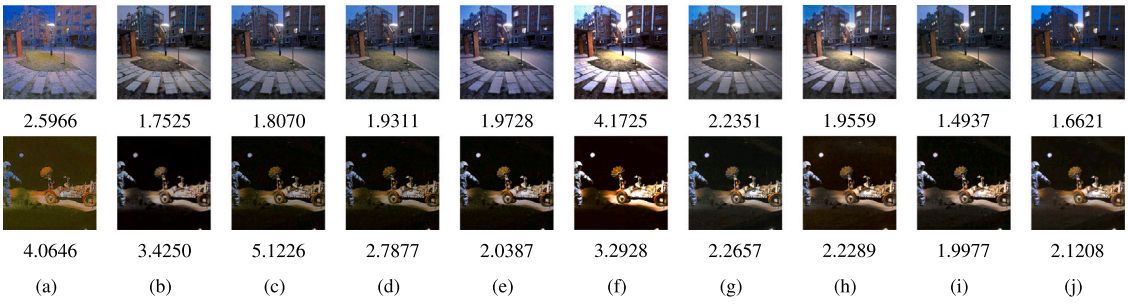



**Fig. 6.** Visual quality comparison of different methods on unpaired datasets. The value under each image denotes its PI (Perceptual Index) value. The smaller the PI value, the better the performance. (a) Retinex [3]; (b) KinD [6]; (c) Zero-DCE [7]; (d) Zero-DCE++ [8]; (e) EnlightenGAN [4]; (f) RUAS [5]; (g) PyNET [15]; (h) Uformer [29]; (i) Ours; (j) Ours++.

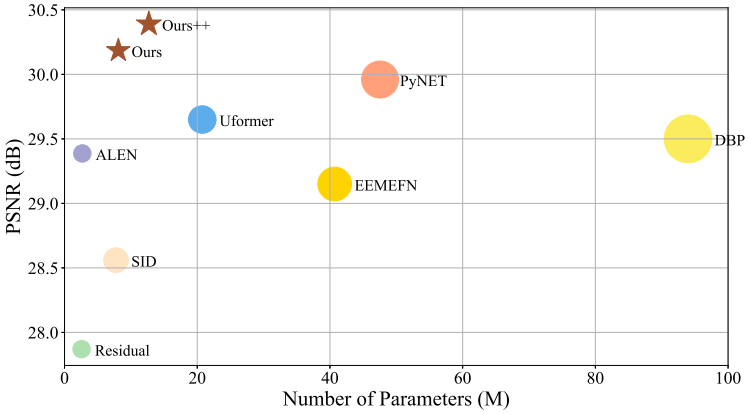



**Fig. 7.** Parameters and PSNR performance comparison of different methods.

is accordingly limited. Constrained by our implementation platform, the GMACs of EEMEFN is not available. Because the proposed cross-scale framework operates on full-resolution images, our GMACs cost reaches as high as 601.9, but is still much smaller than the 1608.68 of the DBP method and 1789.89 of the PyNET. In Fig. 7, we provide a joint comparison of parameters and performance. It is clear that our and ours++ achieve impressive performance with relatively fewer parameters.

Notice that we compare the computational complexity with RAW enhancement models only for a fair comparison. The RGB enhancement models cost much less complexity than these RAW-based models. This is because the RAW enhancement models essentially conduct the ISP function together with the enhancement task, which is much more expensive than the RGB models for the enhancement task only.

### *4.3. Qualitative comparison*

**Comparison on the SID Dataset.** Furthermore, we visualize the enhancement results of all methods for comparison. First, we show the visualization results on the SID dataset which includes the Sony sub-dataset and the Fuji sub-dataset. Observing Fig. 8, one can see that the SID, Residual, EEMEFN, ALEN, DBP, PyNET, and Uformer methods yield poor reconstruction quality, such as obvious blurring artifacts and color bias. The SID and Residual models produce clear pigmentation and artifacts, while the ALEN has less color deviation but still suffers from severe blurring. In terms of local details, the lines produced by DBP are not soft enough and contain some distortions. Taking the bicycle in the fourth line for example, near the bike in the figure, the Residual model produces distinct grid-like streaks and dispersion in the white light; in comparison, the EEMEFN performs significantly better, but the overall tone is warm and deviates a lot from the ground truth; results of the ALEN, PyNET, and DBP are even better, except for the ringing effect in ALEN and PyNET and the block effect in DBP; Similarly, the Uformer results also look blurry; instead, our proposed method achieves a very good exposure effect and improves the image sharpness and contrast. Overall, images enhanced by the proposed method surpass these methods in terms of both color restoration and texture preservation, looking closer to the ground truth. For example, for the last line images, ours can recover “Cardboard” a clearer and more visible font outline whereas visualizations of other images look either blurry or over-smoothed.

**Comparison on the Non-reference Dataset.** We additionally compare different methods on the non-reference dataset, as visualized in Fig. 6. Compared methods include Retinex, KinD, Zero-DCE, Zero-DCE++, EnlightenGAN, RUAS, PyNET, and Uformer. Along with the visualized images, we report the Perceptual Index (PI) value which is based on no-reference image quality measures.

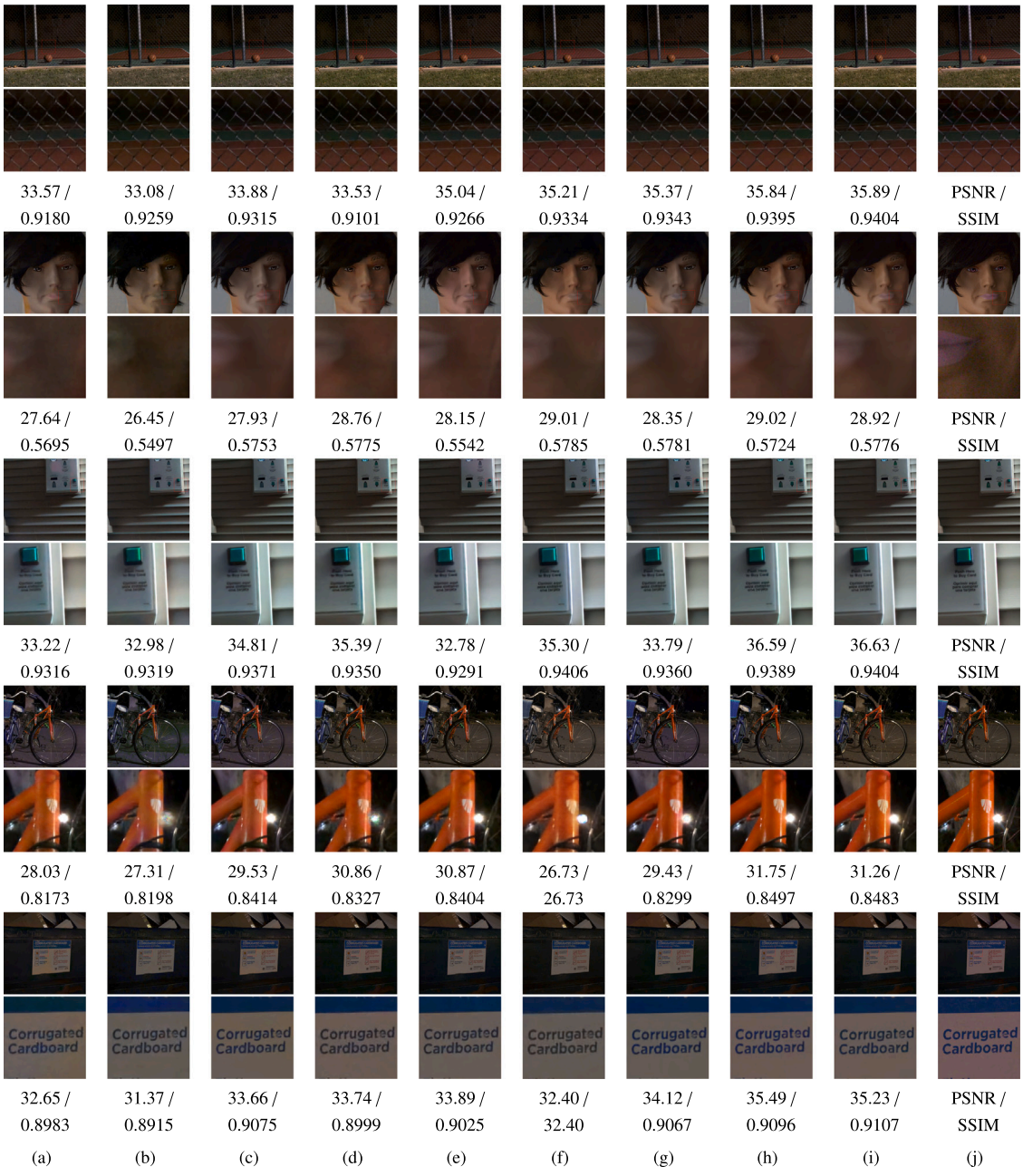



**Fig. 8.** Qualitative comparison of the results of different methods tested on the SID (Sony, the top three lines) and SID (Fuji, the bottom two lines) datasets. (a) SID [9]; (b) Residual [12]; (c) EEMEFN [10]; (d) ALEN [11]; (e) DBP [13]; (f) PyNET [15]; (g) Uformer [29]; (h) Ours; (i) Ours++; (j) Ground truth.

A lower perceptual index indicates better perceptual quality. As seen, our method achieves the best PI value for the exemplified images. We also observe that the proposed model and the KinD model obtain the best quality of all, without obvious over-exposure or under-exposure phenomenon. Relatively, our visualization results look brighter and more natural than that of KinD.

All results above demonstrate that the proposed method attains the best or the second-best results on all metrics, on both RAW and RGB datasets, which is a strong testament to the superiority and generalization of our method.

### *4.4. Ablation study*

In this section, we devise a series of ablation studies to evaluate the contribution of each module in the proposed framework.

**Cross-scale Architecture.** Table 4 reports the enhancement performance using the single main scale, two scales, and three scales of the proposed framework on the SID (Sony) dataset. Using the main scale can provide slightly lower performance to EEMEFN.

**Table 4**
Ablation study on the contribution of different scales in the cross-scale framework.

| Structure | PSNR↑ | MPSNR↑ | SSIM↑ | LPIPS↓ | NIMA↑ |
|---|---|---|---|---|---|
| One scale | 28.9174 | 29.4139 | 0.7750 | 0.4930 | 3.983 ± (1.735) |
| Two scales | 29.9460 | 30.4384 | **0.7875** | 0.4454 | 4.064 ± (1.747) |
| Ours | **30.1851** | **30.6853** | 0.7837 | **0.4424** | **4.092 ± (1.751)** |

**Table 5**
Ablation study on the structure of Transformer and FFC.

| Method | PSNR↑ | MPSNR↑ | SSIM↑ | LPIPS↓ | NIMA↑ |
|---|---|---|---|---|---|
| W/o Trans. & FFC | 28.4132 | 29.1514 | 0.7723 | 0.4677 | 4.071 ± (1.746) |
| W/o Trans. | 29.5845 | 30.0441 | 0.7792 | 0.4429 | 4.062 ± (1.747) |
| W/o FFC | 29.7923 | 30.2181 | 0.7822 | 0.4535 | 4.086 ± (1.749) |
| Only local in FFC | 30.0182 | 30.5094 | 0.7879 | 0.4421 | 4.081 ± (1.749) |
| Only global in FFC | 30.0597 | 30.5997 | **0.7889** | **0.4351** | 4.091 ± (1.750) |
| W/conv fusion in FFC | 30.0759 | 30.5975 | 0.7884 | 0.4380 | **4.095 ± (1.750)** |
| W/o local in ST | 30.0854 | 30.6122 | 0.7843 | 0.4401 | 4.090 ± (1.750) |
| W/o global in ST | 29.0337 | 29.6054 | 0.7781 | 0.4462 | 4.080 ± (1.748) |
| Ours | **30.1851** | **30.6853** | 0.7837 | 0.4424 | 4.092 ± (1.751) |

**Table 6**
Ablation study on the number of Transformers and FFCs.

| | PSNR↑ | MPSNR↑ | SSIM↑ | LPIPS↓ | NIMA↑ |
|---|---|---|---|---|---|
| One FFC | 29.9562 | 30.5089 | 0.7842 | 0.4377 | 4.107 ± (1.751) |
| Three FFC | 29.5714 | 30.5391 | 0.7867 | 0.4402 | **4.123 ± (1.760)** |
| Four FFCs | 29.9901 | 30.5263 | 0.7832 | 0.4477 | 4.089 ± (1.750) |
| One Trans. | 30.0679 | 30.5741 | 0.7835 | 0.4386 | 4.084 ± (1.750) |
| Three Trans. | 29.9674 | 30.5064 | **0.7878** | **0.4358** | 4.089 ± (1.750) |
| Four Trans. | **30.2163** | **30.6925** | 0.7848 | 0.4382 | 4.092 ± (1.750) |
| Ours | 30.1851 | 30.6853 | 0.7837 | 0.4424 | 4.092 ± (1.751) |

Adding the second scale and the third scale, i.e., using more high-level features for global perception brings in significant gains. This also reveals that only local features are insufficient for the whole image enhancement.

**Structure of FFC and Transformer.** To evaluate the performance of the FFC and the Transformer modules in the proposed framework, we replace the FFC and the Transformer using conventional convolutions for performance evaluation, with the results shown in Table 5. As seen, FFC and Transformer bring significant improvement, with the combined use increasing PSNR values by 1.77 dB and separate use by 0.39 dB and 0.60 dB, respectively. In Table 5, w/o Transformer means we replace all Transformers with FFCs; W/o FFC indicates that conventional convolutions are used instead of FFC. It is also found that w/o FFC (using Transformer only) performs slightly better than w/o Transformer, indicating the self-attention mechanism can better exploit pixel correlations in spatial space. Nonetheless, the proposed method achieves the best effect by using paired FFC and Transformer.

We then verify the role of local and global branches in the FFC and the Spectral Transform modules. It can be seen that the local and global branches bring 0.16 dB and 0.12 dB PSNR gains to the FFC, respectively, proving that combining local and global information in FFC is highly effective. Moreover, the combining way of local and global branches also has an impact on the results, for example, using a $3 \times 3$ convolution to fuse the two branches will decrease the PSNR by 0.10 dB but improve the SSIM and NIMA. In addition, it can be seen that disabling the local branch in the Spectral Transform (termed w/o local ST) will cause 0.10 dB PSNR decrease while disabling the global branch (termed w/o global ST) will cause as large as 1.15 dB PSNR and certain SSIM, LPIPS, and NIMA loss.

**The Number of Transformers and FFCs.** One issue well worth considering is the number of Transformers and FFCs, and the depth of the proposed network. To investigate this issue, we decrease or increase the number of FFCs (or Transformers) in the proposed framework for performance evaluation. Notice when we change the number of FFCs (or Transformers), we keep the number of Transformers (or FFCs) constant for fair comparisons.

As shown in Table 6, for each FFCRes Block in Fig. 2, if one FFC, three FFCs, or four FFCs are embedded, the PSNR performance decreases relative to the proposed method with a slight improvement in SSIM, LPIPS, and NIMA metrics. Using more Transformers, e.g., four in this experiment, indeed attains better gains than using two Transformers in the proposed model, e.g., slightly better PSNR (0.03 dB), SSIM, and LPIPS metrics. Similarly, the proposed model achieves better gains than the model using only one Transformer. Although using more Transformer improves the performance slightly, the complexity is increased because of the inherent matrix operations in Transformer. To this end, the proposed method adopts the use of two Transformers.

**Loss Functions.** We evaluate the contribution of each item in our loss function, as described in Table 7. The MAE loss achieves backbone performance. Adding the SSIM and GDL loss gradually improves the PSNR performance. At last, incorporating the FFT loss increases the gain to the best PSNR. The SSIM and the LPIPS are not the best when using the combined loss function. For example,

**Table 7**
Ablation study on the loss function.

| | PSNR↑ | MPSNR↑ | SSIM↑ | LPIPS↓ | NIMA↑ |
|---|---|---|---|---|---|
| $\mathcal{L}_{MAE}$ | 29.7065 | 30.2173 | 0.7834 | 0.4432 | 4.022 ± (1.740) |
| $\mathcal{L}_{MAE} + \mathcal{L}_{SSIM}$ | 29.8795 | 30.4126 | **0.7861** | 0.4354 | 4.073 ± (1.748) |
| $\mathcal{L}_{MAE} + \mathcal{L}_{SSIM} + \mathcal{L}_{GDL}$ | 29.9487 | 30.4551 | 0.7834 | **0.4353** | 4.082 ± (1.750) |
| Ours ($\mathcal{L}_{MAE} + \mathcal{L}_{SSIM} + \mathcal{L}_{GDL} + \mathcal{L}_{FFT}$) | **30.1851** | **30.6853** | 0.7837 | 0.4424 | **4.092 ± (1.751)** |

the best SSIM values are obtained when using MAE loss and SSIM loss only. In general, the proposed loss function attains a balance across all metrics.

## 5. Conclusion

This paper develops a cross-scale framework using paired FFC and Transformer for low-light image enhancement. The cross-scale framework gradually downsamples the input low-light RAW image to three resolutions for low-level, mid-level, and high-level feature extraction. As such, the network captures global and local features which contribute significantly to the enhancement. Within each scale, we additionally introduce the paired FFC and Transformer for spatial–spectral information aggregation. Specifically, features are transformed into the Fourier domain for convolution, enlarging the receptive field; immediately a Transformer is followed to emphasize the important feature and attenuate the useless ones through the self-attention mechanism. Extensive experimental quantitative and qualitative results confirm the effectiveness of the proposed approach. Ablation studies further validate the contribution of each component in our framework.

## CRediT authorship contribution statement

**Yichi Zhang:** Develop the methodology, Create the model, Program. **Hengyu Liu:** Verify and crosscheck results/experiments, Test results and crosscheck research outputs, Data curation. **Dandan Ding:** Write – review & editing, Develop the methodology, Supervise the research activity.

## Declaration of competing interest

No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.compeleceng.2023.108608.

## Data availability

No data was used for the research described in the article.

## Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant 62171174.

## References

[1] Ibrahim H, Pik Kong NS. Brightness preserving dynamic histogram equalization for image contrast enhancement. IEEE Trans Consum Electron 2007;53(4):1752–8. http://dx.doi.org/10.1109/TCE.2007.4429280.
[2] Wang S, Zheng J, Hu H-M, Li B. Naturalness preserved enhancement algorithm for non-uniform illumination images. IEEE Trans Image Process 2013;22(9):3538–48. http://dx.doi.org/10.1109/TIP.2013.2261309.
[3] Wei C, Wang W, Yang W, Liu J. Deep retinex decomposition for low-light enhancement. 2018, arXiv preprint arXiv:1808.04560.
[4] Jiang Y, Gong X, Liu D, Cheng Y, Fang C, Shen X, Yang J, Zhou P, Wang Z. Enlightengan: Deep light enhancement without paired supervision. IEEE Trans Image Process 2021;30:2340–9.
[5] Risheng L, Long M, Jiaao Z, Xin F, Zhongxuan L. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2021.
[6] Zhang Y, Zhang J, Guo X. Kindling the darkness: A practical low-light image enhancer. In: Proceedings of the 27th ACM international conference on multimedia. MM '19, New York, NY, USA: ACM; 2019, p. 1632–40. http://dx.doi.org/10.1145/3343031.3350926, URL http://doi.acm.org/10.1145/3343031.3350926.
[7] Guo CG, Li C, Guo J, Loy CC, Hou J, Kwong S, Cong R. Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2020, p. 1780–9.
[8] Li C, Guo CG, Loy CC. Learning to enhance low-light image via zero-reference deep curve estimation. In: IEEE transactions on pattern analysis and machine intelligence. 2021, http://dx.doi.org/10.1109/TPAMI.2021.3063604.
[9] Chen C, Chen Q, Xu J, Koltun V. Learning to see in the dark. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 2018.
[10] Zhu M, Pan P, Chen W, Yang Y. EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network. Proc AAAI Conf Artif Intell 2020;34(07):13106–13. http://dx.doi.org/10.1609/aaai.v34i07.7013, URL https://ojs.aaai.org/index.php/AAAI/article/view/7013.

[11] Zhang C, Yan Q, Zhu Y, Li X, Sun J, Zhang Y. Attention-based network for low-light image enhancement. In: 2020 IEEE international conference on multimedia and expo (ICME). 2020, p. 1–6. http://dx.doi.org/10.1109/ICME46284.2020.9102774.
[12] Maharjan P, Li L, Li Z, Xu N, Ma C, Li Y. Improving extreme low-light image denoising via residual learning. In: 2019 IEEE international conference on multimedia and expo (ICME). 2019, p. 916–21. http://dx.doi.org/10.1109/ICME.2019.00162.
[13] Karadeniz AS, Erdem E, Erdem A. Burst photography for learning to enhance extremely dark images. IEEE Trans Image Process 2021;30:9372–85.
[14] Sumner R. Processing raw images in matlab. Department of Electrical Engineering, University of California Sata Cruz; 2014.
[15] Ignatov A, Gool L, Timofte R. Replacing mobile camera ISP with a single deep learning model. 2020, p. 2275–85. http://dx.doi.org/10.1109/CVPRW50498.2020.00276.
[16] Kim B-H, Song J, Ye JC, Baek J. PyNET-CA: enhanced PyNET with channel attention for end-to-end mobile image signal processing. In: European conference on computer vision. Springer; 2020, p. 202–12.
[17] Dai L, Liu X, Li C, Chen J. AWNet: Attentive wavelet network for image ISP. In: ECCV workshops. 2020.
[18] Chi L, Jiang B, Mu Y. Fast fourier convolution. Adv Neural Inf Process Syst 2020;33:4479–88.
[19] Zhang Y, Liu H, Ding D, Ma Z. Low light RAW image enhancement using paired fast Fourier convolution and transformer. In: IEEE visual communications and image processing (VCIP). IEEE; 2022.
[20] Li C, Guo C, Han L, Jiang J, Cheng M-M, Gu J, Loy CC. Low-light image and video enhancement using deep learning: A survey. IEEE Trans Pattern Anal Mach Intell 2021.
[21] Ren W, Liu S, Ma L, Xu Q, Xu X, Cao X, Du J, Yang M-H. Low-light image enhancement via a deep hybrid network. IEEE Trans Image Process 2019;28(9):4364–75. http://dx.doi.org/10.1109/TIP.2019.2910412.
[22] Fan M, Wang W, Yang W, Liu J. Integrating semantic segmentation and retinex model for low-light image enhancement. In: Proceedings of the 28th ACM international conference on multimedia. New York, NY, USA: Association for Computing Machinery; 2020, p. 2317–25.
[23] Huang H, Yang W, Hu Y, Liu J, Duan L-Y. Towards low light enhancement with RAW images. IEEE Trans Image Process 2022;31:1391–405. http://dx.doi.org/10.1109/TIP.2022.3140610.
[24] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H, Shao L. CycleISP: Real image restoration via improved data synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2020.
[25] Schwartz E, Giryes R, Bronstein AM. DeepISP: Toward learning an end-to-end image processing pipeline. IEEE Trans Image Process 2019;28(2):912–23. http://dx.doi.org/10.1109/TIP.2018.2872858.
[26] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, http://dx.doi.org/10.48550/ARXIV.2010.11929, URL https://arxiv.org/abs/2010.11929.
[27] Lu Z, Li J, Liu H, Huang C, Zhang L, Zeng T. Transformer for single image super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 457–66.
[28] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H. Restormer: Efficient transformer for high-resolution image restoration. In: CVPR. 2022.
[29] Wang Z, Cun X, Bao J, Liu J. Uformer: A general U-shaped transformer for image restoration. 2021, arXiv preprint arXiv:2106.03106.

**Yichi Zhang** is pursuing the bachelor's degree at the School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China. His research interests include video coding and machine learning.

**Hengyu Liu** is pursuing the bachelor's degree at the School of Information Science and Technology, Hangzhou Normal University, Hangzhou, China. His research interests include image and video processing.

**Dandan Ding** is with the School of Information Science and Technology, Hangzhou Normal University. Her research interests include artificial intelligence based image/video processing, video coding algorithm design and optimization, and point cloud processing.