Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.057662

ARTICLE

Retinexformer+: Retinex-Based Dual-Channel Transformer for Low-Light Image Enhancement

Song Liu[1,2], Hongying Zhang[1,*], Xue Li[1] and Xi Yang[1,3]

[1]School of Information Engineering, Southwest University of Science and Technology, Mianyang, 621000, China

[2]Criminal Investigation Department, Sichuan Police College, Luzhou, 646000, China

[3]School of Electronics and Information, Mianyang Polytechnic, Mianyang, 621000, China

*Corresponding Author: Hongying Zhang. Email: zhanghongying@swust.edu.cn

Received: 24 August 2024 Accepted: 30 October 2024 Published: 17 February 2025

ABSTRACT

Enhancing low-light images with color distortion and uneven multi-light source distribution presents challenges. Most advanced methods for low-light image enhancement are based on the Retinex model using deep learning. Retinexformer introduces channel self-attention mechanisms in the IG-MSA. However, it fails to effectively capture long-range spatial dependencies, leaving room for improvement. Based on the Retinexformer deep learning framework, we designed the Retinexformer+ network. The "+" signifies our advancements in extracting long-range spatial dependencies. We introduced multi-scale dilated convolutions in illumination estimation to expand the receptive field. These convolutions effectively capture the weakening semantic dependency between pixels as distance increases. In illumination restoration, we used Unet++ with multi-level skip connections to better integrate semantic information at different scales. The designed Illumination Fusion Dual Self-Attention (IF-DSA) module embeds multi-scale dilated convolutions to achieve spatial self-attention. This module captures long-range spatial semantic relationships within acceptable computational complexity. Experimental results on the Low-Light (LOL) dataset show that Retexformer+ outperforms other State-Of-The-Art (SOTA) methods in both quantitative and qualitative evaluations, with the computational complexity increased to an acceptable 51.63 G FLOPS. On the LOL_v1 dataset, RetinexFormer+ shows an increase of 1.15 in Peak Signal-to-Noise Ratio (PSNR) and a decrease of 0.39 in Root Mean Square Error (RMSE). On the LOL_v2_real dataset, the PSNR increases by 0.42 and the RMSE decreases by 0.18. Experimental results on the Exdark dataset show that Retexformer+ can effectively enhance real-scene images and maintain their semantic information.

KEYWORDS

Low-light image enhancement; Retinex; transformer model

1 Introduction

Brightness information is a key indicator of image content representation. Enhancing the brightness of low-light images is an important research direction in the field of computer graphics.

Traditional image processing methods for low-light enhancement include histogram equalization [1–5] and gamma correction [6–8]. These methods are based on fundamental image principles and

are highly interpretable. However, their applicability is limited in complex lighting scenes. According to visual imaging principles, light illuminates different object surfaces, and objects with different materials have distinct reflective properties. These reflected rays ultimately project onto the retina to form an image. The Retinex theory [9] decomposes an image into illumination and reflection components. It enhances low-light images by adjusting the illumination component. This approach provides theoretical support for addressing challenging low-light enhancement problems.

With the development of artificial intelligence technology, especially the wide application of deep learning methods in the field of image processing, new ideas and methods have been provided for low-light image enhancement. Recent research has applied convolutional neural networks [10–13] and transformer models [14–18] to low-light image enhancement. This research has achieved significant progress. Convolutional networks can effectively capture the regional spatial contextual information of images. They have shown certain efficacy in low-light image enhancement. However, convolutional networks have limitations in capturing long-range dependencies. Transformer-based methods express spatial long-range dependencies by introducing self-attention mechanisms. These methods can better restore lighting details in low-light images. The computational complexity of transformers is typically proportional to the square of the spatial size. This results in slow inference speed and high computational resource consumption.

Based on the summary of the Related Work, the Materials and Methods section provides a detailed description of the principles and framework of the proposed method. The framework consists of two parts: illumination estimation and damage restoration. The illumination estimator utilizes a multi-scale expanded convolution module to extract spatially distant semantic information. It represents the illumination characteristics of different regions and generates a perturbed illumination map. The illumination fusion module (IFU) of the damage restorer adopts the Unet++ network structure. This algorithm incorporates multi-level skip connections to better fuse semantic information from different levels and reduce information loss. The illumination fusion attention block (IFAB) introduces a spatial attention mechanism to express spatial distant dependencies within an acceptable computational complexity. The key branch of spatial attention employs multi-scale unfold convolutions to extract spatial position features. In the "Results" section, we first evaluate and validate the advancements of our method in image enhancement from both quantitative and qualitative dimensions using the Low-Light (LOL) dataset. We also compare the computational complexity and parameter count. Next, we demonstrate the effectiveness of our method in enhancing real-scene images using the Exdark dataset and conduct object detection experiments using recommended weights from YOLOv3 to evaluate the preservation of semantic information in enhanced images. In the ablation study, we compare the gains brought by different improvement details to the enhancement effect to support the effectiveness of our method. Finally, we summarize the strengths and weaknesses of our method and propose future research directions.

Extensive experiments show that Retinexformer+ achieves better quantitative and qualitative results on the LOL dataset [19,20]. It outperforms other state-of-the-art (SOTA) supervised and unsupervised methods [21–25]. Our main innovations are as follows:

1. Proposing a multi-scale dilated convolution structure. This structure expands the receptive field while expressing the weakening semantic dependencies between pixels as the distance increases.

2. Adopting Unet++ multi-level skip connections in damage restoration. This approach better fuses semantic information from different levels and reduces information loss.

3. Designing a novel multi-scale dilated convolution spatial attention module. This module expresses spatial long-range semantic relationships. It reduces the computational complexity of transformer spatial attention from quadratic to linear with respect to spatial size.

2 Related Work

2.1 Classical Image Processing Methods

Classical image processing methods include histogram equalization [3–5,26,27] and gamma correction [6–8]. Cheng et al. [3] proposed a multi-peak generalized histogram equalization method that improves global histogram equalization by using multi-peak histogram equalization combined with local information. Lee et al. [4] proposed a novel contrast enhancement algorithm based on the layered difference representation of 2D histograms to enhance image contrast by amplifying the gray-level differences between adjacent pixels. Wang et al. [6] proposed a new method combining dynamic contrast ratio enhancement and inverse gamma correction for alternating current plasma display panel (AC PDP), and both are realized simultaneously. Huang et al. [7] proposed an automatic transformation technique is presented which improves the brightness of dimmed images via gamma correction and luminance pixel probability distribution and uses temporal information for video enhancement to reduce computational complexity. Rahman et al. [8] proposed an adaptive gamma correction method where parameters are set dynamically based on image information to appropriately enhance the contrast of the image. These methods enhance images by adjusting brightness and contrast. These methods are simple and highly interpretable. However, they do not account for complex real-world lighting scenarios. As a result, enhanced images often lack naturalness and realism.

2.2 Traditional Cognition Methods

According to the Retinex theory [9], light illuminates different object surfaces, and objects with different materials have distinct reflective properties. These reflections form images on the retina. The Retinex theory decomposes an image into illumination and reflectance components. It enhances low-light images by adjusting the illumination component [28–32]. Fu et al. [28] proposed a fusion-based method for enhancing weakly illuminated images using multiple techniques, including decomposing, deriving, designing, fusing, and compensating to obtain an enhanced image for different weak illumination conditions. Fu et al. [29] proposed a weighted variational model to estimate the reflectance and illumination from an observed image. Guo et al. [30] proposed a simple yet effective low-light image enhancement (LIME) method. Wang et al. [32] made three major contributions including proposing a lightness-order-error measure, a bright-pass filter for image decomposition and a bi-log transformation for mapping illumination. However, these methods require manual setting of illumination priors. Inaccurate priors can lead to artifacts and color distortions in the enhanced results. Additionally, traditional methods often neglect the impact of noise. Simple brightness enhancement can retain and amplify noise.

2.3 Deep Learning Methods

In 2017, Lore et al. [33] proposed a deep learning-based low-light image enhancement algorithm. Convolutional neural network (CNN) methods [33–37] have been widely applied in low-light image enhancement. Hong et al. [34] proposed a novel unsupervised low-light image enhancement network named LE-GAN based on generative adversarial networks and trained with unpaired low-light images. Lore et al. [33] proposed a deep autoencoder-based approach to identify signal features from low-light images and adaptively brighten images without over-amplifying lighter parts in high dynamic range

images. Sharma et al. [35] proposed a method involving estimating the camera response function, decomposing the linearized image into LF and HF feature maps, processing them separately, and combining them to generate an output with increased dynamic range and suppressed light effects. Jiang et al. [36] introduced a highly effective unsupervised generative adversarial network that can be trained without low-light image pairs and generalizes well on real-world test images. Wei et al. [38] and subsequent studies [20,39] combined deep learning methods with the Retinex theory. They adjusted and optimized the Retinex model parameters through convolutional networks. These studies showed improvements over traditional cognitive methods. However, these methods often employ multi-stage training processes. This can be cumbersome. In 2019, Wang et al. [12] proposed a single-stage CNN method to directly predict illumination maps. However, this approach can cause color distortions and noise amplification while enhancing the brightness of low-light images. CNN-based methods still have limitations in expressing long-range spatial dependencies.

2.4 Vision Transformer

In 2017, Vaswani et al. [40] presented the Transformer, a neural network architecture relying on self-attention to process sequential data rather than traditional recurrent or convolutional layers. Vision Transformer converts images into sequence data by splitting them into patches and mapping each patch to a vector for processing. Recently, Vision Transformers and their variants have been applied to low-light image enhancement. In 2022, Xu et al. [41] proposed SNR-Net, a CNN-Transformer hybrid model for low-light image enhancement. Due to the computational complexity of Transformers being proportional to the square of the spatial size, SNR-Net integrates a single global Transformer only at the lowest resolution of the U-shaped network. In 2023, Retinexformer [14], based on the Retinex theory, designed a single-stage Transformer network. It concatenates the height and width of the image into HW tokens and performs self-attention calculations in the channel direction. Spatial position information is extracted through two layers of 3×3 convolutional networks. Retinexformer achieved SOTA results that year. However, there is still room for improvement in extracting long-range spatial dependency features. The application of deep learning methods has significantly improved low-light image enhancement. However, further research is needed to enhance the extraction of long-range spatial dependency features under acceptable computational complexity.

3 Materials and Methods

Fig. 1 shows the framework of Retinexformer+. Retinexformer+ consists of an illumination estimator (Fig. 1a) and an illumination-fused U-Net (IFU) (Fig. 1b). The illumination estimator uses multi-scale dilated convolution. This method extracts image features and expresses pixel semantic dependencies. The damage restorer is designed as an illumination multi-scale fusion U-Net (IFU). Each layer of the IFU uses an illumination-fused attention block (IFAB) to fuse features (Fig. 1b).

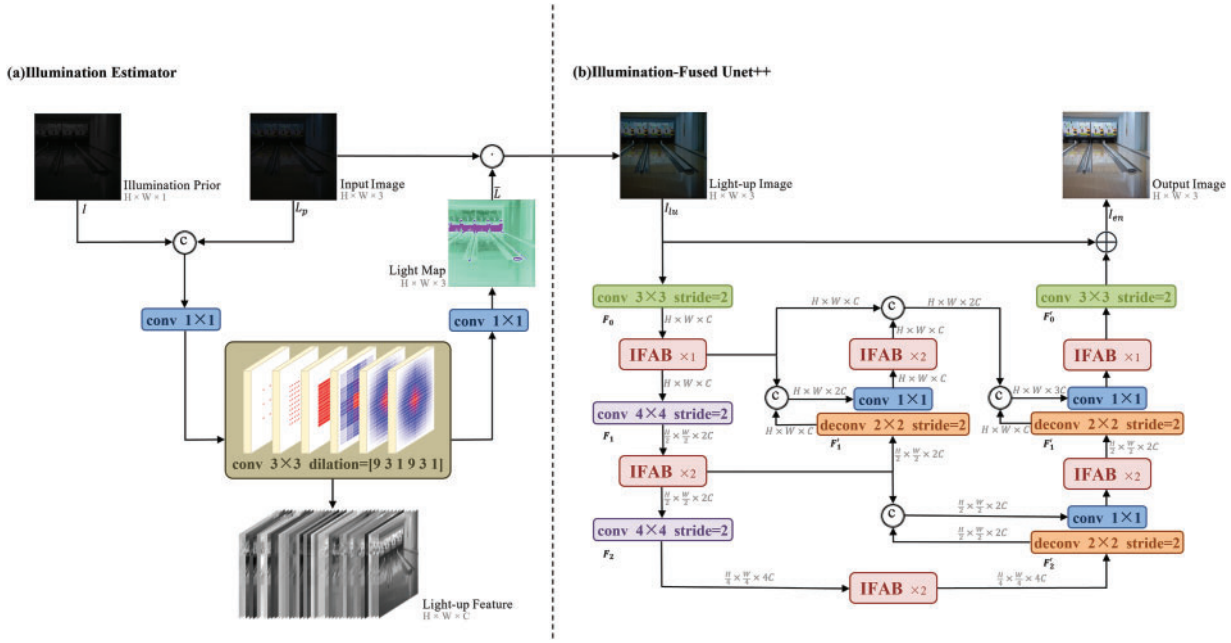

Figure 1: Retinexformer+ consists of an (a) illumination estimator and an (b) illumination-fused U-Net

3.1 Network Framework Design

The Retinex image enhancement algorithm decomposes a low-light image $I \in R^{H\times W\times 3}$. It splits it into a reflection component $R \in R^{H\times W\times 3}$ and an illumination component $L \in L^{H\times W}$, as follows:

$$I = R \odot L \tag{1}$$

The reflection component R expresses the inherent reflective properties of objects. The illumination component L describes the distribution of light in the scene. In real low-light scenarios, image acquisition often includes noise and artifacts. Reflection disturbances are expressed as $\hat{R} \in R^{H\times W\times 3}$. The illumination distribution is uneven, with weaker illumination in shadowed areas or multiple light sources in low-light environments. Illumination disturbances are expressed as $\hat{L} \in L^{H\times W}$. Considering the multi-source light field in real scenes, the illumination component can be decomposed as $L = \sum_{i=1}^{N} L_i$. The illumination disturbance can be decomposed as $\hat{L} = \sum_{i=1}^{N} \hat{L}_i$. By introducing reflection disturbances and multi-source light field disturbances, the low-light enhancement function can be expressed as:

$$I = \left(\sum_{i=1}^{N} L_i + \sum_{i=1}^{N} \hat{L}_i\right) \odot \left(R + \hat{R}\right) = \sum_{i=1}^{N} L_i \odot R + \sum_{i=1}^{N} \hat{L}_i \odot R + \left(\sum_{i=1}^{N} L_i + \sum_{i=1}^{N} \hat{L}_i\right) \odot \hat{R} \tag{2}$$

Multiplying both sides of the equation by the illumination map $\overline{L} = \left(\sum_{i=1}^{N} L_i\right)^{-1}$, we get:

$$\overline{L} \odot I = R + \sum_{i=1}^{N} \hat{L}_i \odot \overline{L} \odot R + \left(\sum_{i=1}^{N} L_i + \sum_{i=1}^{N} \hat{L}_i\right) \odot \overline{L} \odot \hat{R} \tag{3}$$

In this equation, $\overline{L} \odot I$ is the illuminated image, denoted as I_{lu}. R is the normally exposed image. $\sum_{i=1}^{N} \hat{L}_i \odot \overline{L} \odot R$ is the color distortion part after illumination. $\left(\sum_{i=1}^{N} L_i + \sum_{i=1}^{N} \hat{L}_i\right) \odot \overline{L} \odot \hat{R}$ is the noise and artifacts enhanced during illumination.

3.2 Illumination Estimator

In the implementation of [14], each pixel of the input image I is averaged along the channel dimension to obtain the illumination prior L_p. The input image I and L_p are concatenated into a four-channel input. A 5×5 convolution outputs the 40-channel light-up features L_{lu}. A 1×1 convolution converts the 40 channels into three channels, generating the light-up map $\overline{L} = \left(\sum_{i=1}^{N} L_i\right)^{-1}$. The 40-channel light-up features F_{lu} describe various illumination characteristics. However, a 5×5 convolution kernel is insufficient to extract the spatial distribution features of multiple light sources.

To better extract multi-source distribution features, the convolution network design should consider two aspects. These are expressing long-range spatial semantic relationships and expressing semantic weights between pixels. The number of convolution layers and the convolution kernel size are crucial to achieving these goals. Multi-scale dilated convolution, with different dilation coefficients, can expand the receptive field. This method controls computational complexity and captures multi-scale contextual information. The HDC [37] rule should be followed in designing dilated convolutions. This ensures continuous coverage of the focus area and avoids holes or missing edges. The HDC rules are as follows:

1. The maximum distance between two non-zero elements in the second layer should be less than the size of the convolution kernel of that layer.
2. Convolution coefficients should be set in a sawtooth pattern.
3. The greatest common divisor of the dilation coefficients should not exceed one.

To balance computational complexity, the convolution kernel size is set to 3. This follows the HDC rule, where the maximum distance between non-zero elements in the second layer is less than 3. We use six layers of convolution with a sawtooth dilation coefficient pattern r = 9,3,1,9,3,1. This generates a convolution receptive field of size 53×53.

When the dilation coefficient r = 9, as shown in Fig. 2a. When when r = 9,3, as shown in Fig. 2b, the maximum distance between non-zero elements is less than 3. When r = 9,3,1, as shown in Fig. 2c, the convolution is equivalent to a normal convolution with a kernel size of 27×27. When r = 9,3,1,9, as shown in Fig. 2d. When r = 9,3,1,9,3, as shown in Fig. 2e. When r = 9,3,1,9,3,1, as shown in Fig. 2f, the semantic weights between pixels decrease with increasing distance from the central pixel.

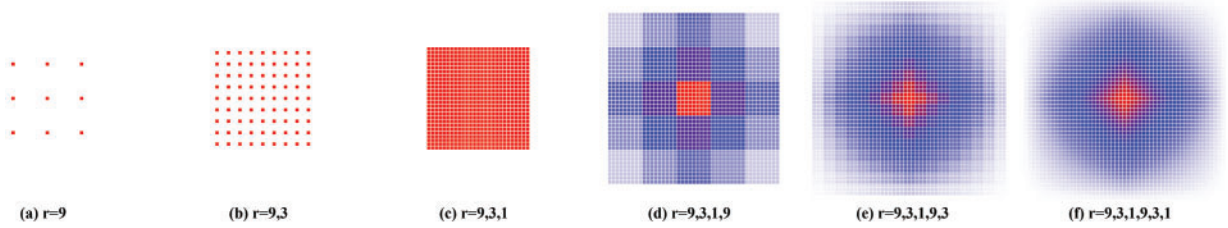

Figure 2: Multi-scale dilated convolution features

3.3 Illumination-Fused Unet++

Eq. (3) shows that spatial weight disturbances of multi-source illumination on the enhanced image cause noise and color distortion. To address this issue, a three-scale Unet++ structure is designed, as shown in Fig. 1b. This structure uses multi-level skip connections. It can extract and fuse more semantic information at different levels. This reduces information loss and improves segmentation accuracy. It better expresses the spatial distribution features of multi-source illumination.

As shown in Fig. 1b, the IFU is designed as a three-layer Unet++ [42] structure. Unet++ is a deeply supervised encoder-decoder network. The encoder and decoder sub-networks are connected through a series of nested and dense skip pathways. Compared with the Unet structure, it has enhanced feature extraction ability and better performance. In the encoding part, the illuminated image Ilu undergoes a 3×3 convolution with a stride of 2, an IFAB encoding, a 4×4 convolution with a stride of 2 for down-sampling, two IFAB encodings, and another 4×4 convolution with a stride of 2 to generate multi-scale encoded features $F_i \in R^{\frac{H}{2^i} \times \frac{W}{2^i} \times 2^i C}$ where i = 0,1,2. Subsequently, F_2 undergoes two IFAB encodings. The decoding part features a symmetric up-sampling branch, with up-sampling using 2×2 deconv after two IFAB encodings of F_1 and F_2, adding skip connections at the same level to reduce information loss during encoding and decoding. Finally, the decoding part outputs the residual image I_{re}, enhancing the illuminated image through $I_{en} = I_{lu} + I_{re}$.

3.4 Illumination-Fused Attention Block

Fig. 3 shows the process flow of obtaining output features by fusing Light-up Feature and Input Feature through the IFAB. Fig. 3a shows the structure of the IFAB module, which consists of a Layer Normalization (LN) layer, an Illumination-Fused Dual Self-Attention (IF-DSA) module, and a Feed-Forward Neural Network (FNN). The structure of the IF-DSA module is shown in Fig. 3b.

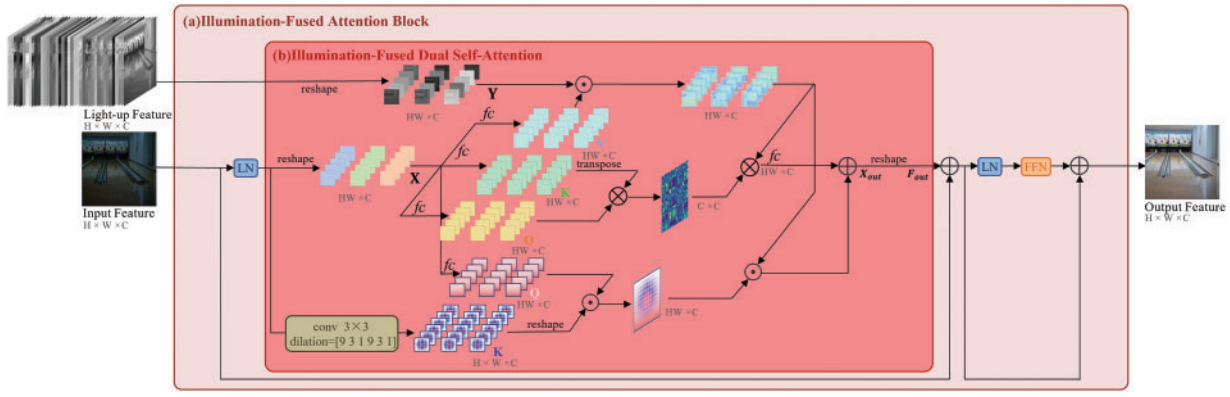

Figure 3: The Illumination-Fused Attention Block (IFAB) consists of a Layer Normalization (LN) layer, an Illumination-Fused Dual Self-Attention (IF-DSA) module, and a Feed-Forward Neural Network (FNN)

3.5 Illumination Fusion with Dual Self-Attention

By observing the input image, it is clear that different regions are illuminated by different light sources. Therefore, it is necessary to extract semantic information from different regions of the image to encode the illumination features and achieve brightness enhancement. Retinexformer [14] reshapes the input features $F_{in} \in R^{H \times W \times C}$ into $X \in R^{HW \times C}$ and uses three bias-free fully connected layers to linearly

map X to $Q = XW_Q^T, K = XW_K^T$ and $V = XW_V^T \in R^{HW\times C}$ (W_Q, W_K, W_V are learnable parameters of the linear mapping layers, and T represents matrix transpose). The input light-up feature F_{lu} is reshaped into $Y \in R^{HW\times C}$, then Q, K, V and Y are each split into k heads:$Q = [Q_1, Q_2, \ldots, Q_k], K = [K_1, K_2, \ldots, K_k], V = [V_1, V_2, \ldots, V_k], Y = [Y_1, Y_2, \ldots, Y_k]$, where $Q_i, K_i, V_i, Y_i \in R^{HW\times d_k}, d_k = \frac{C}{k}$, and $i = 1, 2, \ldots, k$. The channel self-attention for each head is represented as:

$$Channel\_Attention\,(Q_i, K_i, V_i, Y_i) = (Y_i \odot V_i) \odot softmax\left(\frac{K_i^T Q_i}{(\alpha\_c_i)}\right) \tag{4}$$

where $(\alpha\_c_i) \in R^{d_k\times d_k}$ is a learnable adaptive scaling parameter matrix. The spatial context information is extracted using two 3×3 convolutions to obtain the spatial position information $P \in R^{C\times HW}$, which is then connected to the channel multi-head self-attention. Finally, after reconstruction, the output feature $F_{out} \in R^{H\times W\times C}$ is obtained.

Retinexformer [14] cleverly compresses the spatial HW dimensions in the self-attention computation, making the computational complexity of the attention calculation only the square of the single-head channel size. This ensures that the computational complexity remains within an acceptable range while effectively obtaining channel attention to express different light-up features. However, using only two 3×3 convolutions to extract spatial semantic information is insufficient to express long-range dependencies in the space. Directly performing self-attention calculations on the spatial dimensions would result in a computational and memory resource overhead proportional to the square of the spatial size $H \times W$, which is undeniably substantial. Therefore, expressing long-range dependencies in space while controlling computational complexity is a worthwhile research problem.

Retinexformer+ reshapes the input features $F_{in} \in R^{H\times W\times C}$ into $X \in R^{C\times H\times W}$. It uses bias-free fully connected layers to linearly map X to $Q = XW_Q^T$ and $V = XW_V^T$ (W_Q and W_V are learnable parameters of the linear mapping layers, and T represents matrix transpose). After X undergoes 3×3 convolution operations with dilation coefficients r = 9,3,1,9,3,1, $K^{C\times H\times W}$ is obtained. Q and K are concatenated, followed by two layers of 1×1 convolutions for fusion, achieving spatial position self-attention calculation. The input light-up feature F_{lu} is reshaped into $Y \in R^{HW\times C}$, and the spatial self-attention calculation formula is expressed as:

$$Spatial\_Attention\,(Q, K, V, Y) = (Y \odot V) \odot softmax\left(\frac{conv\,(K) \odot Q}{\alpha\_s}\right) \tag{5}$$

where $\alpha\_s \in R^{H\times W}$ is a learnable adaptive scaling parameter matrix. Finally, the channel self-attention in Retinexformer [14] is reconstructed into $Channel\_Attention \in R^{C\times H\times W}$ and connected to the spatial self-attention $Spatial\_Attention \in R^{C\times H\times W}$, ultimately reconstructing the output feature $F_{out} \in R^{H\times W\times C}$.

3.6 Complexity Analysis

The computational complexity of the IF-DSA module includes the complexity of convolution operations and pointwise multiplication operations between matrices. The computational complexity for six convolution operations with a kernel size of 3×3 is $O\,(Conv) = H \times W \times C^2 \times 3^2 \times 6 = 54HWC^2$. The complexity for pointwise multiplication operations between matrices is $O\,(Pointwise\_Multiply) = HWC$. Therefore, the complexity of the IF-DSA module can be expressed as:

$$O\,(IF - DSA) = O\,(Conv) + O\,(Pointwise\_Multiply) = \left(54C^2 + C\right) HW \tag{6}$$

In contrast, the computational complexity of traditional spatial self-attention methods, like the global MSA used in SNR-Net, is:

$$O\left(G-MSA\right)=2C\left(HW\right)^{2} \tag{7}$$

Comparing Eqs. (6) and (7) show that the IF-DSA module expresses long-range semantic relationships in space. It reduces the computational complexity of spatial self-attention methods from the square of the spatial size to a linear function of the spatial size.

4 Results

4.1 Datasets and Implementation Details

We conducted experiments on the LOL dataset. The LOL dataset is divided into LOL_v1 [38] and LOL_v2 [19]. LOL_v1 contains 485 pairs of training data. It contains 15 pairs of testing data. LOL_v2 is further divided into LOL_v2_real and LOL_v2_synthetic. LOL_v2_real consists of real-scene photographs, containing 689 pairs of training data and 100 pairs of testing data. LOL_v2_synthetic consists of synthetic data. It contains 900 pairs of training data and 100 pairs of testing data. Each pair of data in all datasets includes one standard reference image. Each pair also includes one corresponding low-light image.

We implemented the Retinexformer+ model based on PyTorch. Training and testing were conducted on a Linux server equipped with a 3090 24 GB GPU. The system environment included CUDA 11.8, Python 3.7, and PyTorch 1.13. During training, the data size for LOL_v1 and LOL_v2_synthetic was set to 128×128, with a batch size of 8. For LOL_v2_real, the data size was 256×256, with a batch size of 8. We used random rotation and flipping to augment the training data. The training was performed using the Adam optimizer, with momentum terms of 0.9 and control parameters of 0.999. The aim was to minimize the Mean Absolute Error (MAE) between the enhanced image and the ground truth image. The initial learning rate was set to 2×10^{-4}. It was gradually decreased to 1×10^{-6} using a cosine annealing strategy.

4.2 Quantitative Results

In the quantitative evaluation, we used Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Root Mean Square Error (RMSE) as metrics. Higher PSNR indicates better image enhancement. Higher SSIM indicates better retention of high-frequency details and structures. Lower RMSE indicates better model performance. As shown in Table 1, our method improves the PSNR by 1.15 on the LOL_v1 dataset. Our method improves the PSNR by 0.42 on the LOL_v2_real dataset. The RMSE decreases by 0.39 on the LOL_v1 dataset. The RMSE decreases by 0.18 on the LOL_ v2_real dataset.

Table 1 compares the evaluation metrics of various supervised and unsupervised state-of-the-art (SOTA) methods. The comparison is based on the publicly available LOL_v1 and LOL_v2_real datasets. The data in the table were directly quoted from these papers.This demonstrates the superior performance of our method.

Our method has FLOPs and Params that are 3.31 times and 2.20 times larger than Retinexformer, and 1.21 times and 0.77 times larger than RetinexMamba. The computational complexity and parameter size are acceptable.

Table 1: Quantitative comparisons on LOL_v1 [38] and LOL_v2 [19]. (Red data represents the best results, and blue data represents the second-best results. Our Retinexformer+ algorithm significantly outperforms other algorithms)

Methods	Complexity		LOL_v1			LOL_v2_real		
	FLOPS (G)	Params (M)	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Supervised								
MBLLEN [10]	0.45	0.01	17.94	0.70	18.78	15.95	0.70	30.22
Retntinex-Net [38]	587.47	0.84	17.19	0.59	22.59	16.41	0.64	20.21
KinD [39]	34.99	8.02	20.35	0.81	14.30	18.07	0.78	18.04
KinD++ [20]	35.06	8.27	20.71	0.80	14.34	16.80	0.74	15.64
MIRNet [13]	785	31.79	24.14	0.84	12.03	20.36	0.78	14.21
URetntinex-Net [43]	53.02	0.34	21.45	0.80	13.55	21.55	0.80	14.23
Retinexformer [44]	15.57	1.61	23.86	0.83	8.30	21.93	0.84	9.56
RetinexMamba [44]	42.82	4.59	24.03	0.83	8.17	22.45	0.84	9.38
Retinexformer+	51.63	3.55	25.29	0.84	7.78	22.87	0.84	9.20
Unsupervised								
Zero-Dce [22]	4.83	0.08	16.76	0.56	34.42	18.06	0.58	29.01
RUAS [23]	0.83	0.003	16.40	0.50	30.21	16.87	0.51	29.23
SCI [24]	0.02	0.003	14.86	0.54	24.87	15.34	0.52	27.50
PairLie [25]	20.81	0.34	19.69	0.71	19.03	19.29	0.68	20.01
NeRCO [45]	344.53	23.30	19.70	0.77	24.80	19.23	0.67	23.13
CLIP-LIE [46]	16.96	0.28	17.21	0.59	10.18	17.06	0.59	10.64
Enlighten-Your-Voice [47]	0.23	0.001	19.73	0.72	10.13	19.34	0.69	10.21

4.3 Qualitative Results

Figs. 4 and 5 show the qualitative comparison results of Retinexformer+ with other SOTA algorithms on the LOL_v1 and LOL_v2_real datasets, respectively. Upon zooming in, it can be observed that the images processed by the Retinex method exhibit significant noise and artifacts. KinD and Uretinex-Net show overexposure and underexposure in different regions. Retinexformer+ restores the color of the stapler more realistically than the Retinexformer and Retinexmamba methods. Furthermore, to visually demonstrate the effects of our method, we compared the HSV color space images corresponding to the images on the LOL_v1 and LOL_v2_real datasets in Fig. 6. It can be observed that the color distribution of the images processed by our method is closer to the Ground Truth images.

Figure 4: Qualitative results on the LOL_v1 dataset, showing that our method effectively controls exposure intensity across different regions, reduces noise, and achieves color reproduction closest to the ground truth image

Figure 5: Qualitative results on the LOL_v2_real dataset, showing that our method effectively controls exposure intensity across different regions, reduces noise, and achieves color reproduction closest to the ground truth image

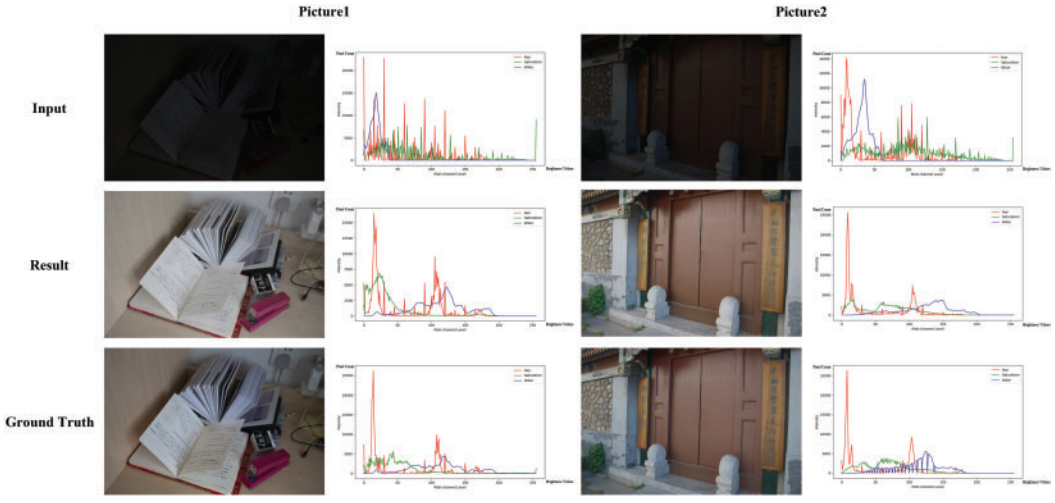

Figure 6: Image display of the HSV color space

4.4 Low-Light Object Detection

In order to evaluate the preservation of semantic information by enhancement methods, we followed the evaluation method in References [48,49]. We conducted brightness enhancement on underexposed images from real scenes. Subsequently, we verified the effects using existing object detection models. The Exdark dataset, containing 7363 underexposed images with annotations for 12 object categories, was chosen for the experiment. After applying the brightness enhancement method described in Table 2, we used the recommended weights of the YOLOv3 algorithm to detect the objects. The COCO dataset provides annotations for 80 different object categories found in real scenes. All 12 object categories from Exdark have corresponding types in COCO. We only needed to change the annotations for "People" and "Table" to "Person" and "Dining Table" in COCO. Finally, we evaluated the performance using the mAP metric.

Table 2: Target detection results of YOLOv3 on the Exdark dataset after enhancement by different algorithms (Red data represents the best results, and blue data represents the second-best results)

Methods	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motor	Person	Dining table	Mean
Original LLIs	67.59	58.09	58.79	77.29	69.99	53.69	43.99	55.89	60.99	56.09	64.09	38.89	58.78
Retinex	56.22	49.42	51.72	70.42	62.42	43.62	36.82	48.22	50.12	47.82	52.62	31.12	50.05
KinD	62.44	56.24	56.04	75.64	67.24	48.64	42.14	53.54	55.54	53.74	60.74	36.24	55.68
Uretinex-Net	67.83	59.53	60.73	80.33	72.43	54.53	46.13	59.13	64.33	58.23	65.83	41.73	60.90
Retinexformer	73.86	63.96	62.66	81.16	72.96	60.66	49.06	59.09	63.59	61.16	65.89	43.06	63.09
Retinexmamba	72.59	63.19	62.59	81.59	75.05	58.85	50.29	60.86	65.59	60.69	63.16	42.89	63.11
Retinexformer+	73.65	64.15	62.75	81.15	75.19	60.19	49.85	61.25	65.76	60.75	66.15	43.25	63.67

The results in Table 2 show that images enhanced by our method achieved the highest average precision: 63.67 AP in the YOLOv3 model. This was a 0.56 AP improvement over the second highest. Furthermore, our method also had the best results in detecting categories such as Boat, Bottle, Car, Cup, Dog, Person, and Dining Table. It is worth noting that the detection accuracy for Retinex and KinD enhanced images did not improve.

Fig. 7 shows target detection results of original and enhanced images. Our method suppresses noise and has generalization for real scene image enhancement in Exdark dataset. Using the recommended weights of YOLOv3 to perform target detection on pictures, original image has high detection confidence but missed detections. Retinex-enhanced image reduces confidence and has false detections. The KinD-enhanced image detects a car on the right side, but confidence decreases. Uretinex-Net misdetects car as person. Retinexformer slightly reduces person detection confidence. Retinexmamba significantly reduces person detection confidence. Our method reduces missed and false detection rates while maintaining confidence.

Figure 7: Comparison of object detection in low-light scenes using different methods on the Exdark dataset

4.5 Ablation Study

We conducted ablation studies on three datasets: LOL_v1, LOL_v2_real, and LOL_v2_synthetic. We set up four different framework models to verify the contribution of each designed component to the improvement of the algorithm's performance.

"Dilated conv" indicates using a six-layer dilated convolution structure with dilation rates of r = 9,3,1,9,3,1 during the illumination estimation phase. This structure expands the receptive field. It also captures how semantic dependencies between pixels weaken with increasing distance.

"Unet++" indicates using Unet++ with multi-level skip connections for damage restoration. This approach better integrates semantic information from different levels and reduces information loss.

"Conv-transformer" indicates adding multi-scale dilated convolution spatial attention to the IF-DSA module. This addition helps express long-range spatial semantic relationships.

"Transformer+" indicates the experimental results after incorporating all components into the network.

Under the same local configuration environment, we tested the PSNR, SSIM, and RMSE values for each framework model. The improvements of each component enhance the algorithm's performance to varying degrees, as shown in Table 3. The data in the table are from the code testing results. Retinexformer+ performs the best in parameter evaluation metrics. This confirms that our network design is reasonable and effective.

Table 3: Ablation Study on LOL_v1, LOL_v2_real and LOL_v2_syn (Red data represents the best results)

Methods	LOL_v1			LOL_v2_real			LOL_v2_syn		
	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓	PSNR↑	SSIM↑	RMSE↓
Retinexformer [45]	23.86	0.83	8.30	21.93	0.84	9.56	25.37	0.93	8.43
Dilated conv	23.95	0.83	8.29	22.09	0.85	9.59	25.88	0.93	8.12
Unet++	24.06	0.83	8.33	22.11	0.83	9.60	25.56	0.93	8.28
Conv-transformer	24.02	0.83	8.34	22.35	0.84	9.37	25.82	0.93	8.13
Retinexformer+	25.29	0.84	7.78	22.87	0.84	9.20	26.07	0.93	8.00

5 Conclusion

This paper introduces Retinexformer+, a new architecture designed for low-light image enhancement. This architecture uses a six-layer dilated convolution to capture long-range spatial semantic features. This approach better represents the distribution characteristics of multiple light sources. The damage restorer utilizes Unet++ with multi-level skip connections. This design effectively integrates semantic information from different levels and reduces information loss. Most importantly, the IF-DSA module includes a novel convolution-based spatial attention module. This module captures long-range spatial dependencies. It reduces the computational complexity of transformer spatial attention. The complexity is reduced from being proportional to the square of the spatial size to being proportional to a multiple of the spatial size.

Based on qualitative and quantitative experimental analyses on the LOL dataset, Retinexformer+ outperforms current state-of-the-art methods. Experiments on the Exdark dataset verify its generalization for enhancing real-scene images. Detection results with YOLOv3 weights verify that Retinexformer+ can retain image semantic information while enhancing low-illumination images. Although Retinexformer+'s computational complexity is acceptable, it increases module parameters and consumes more resources. Future work will focus on improving model performance and reducing resource consumption, called "Retinexformer++".

Acknowledgement: The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

Funding Statement: This work was supported by the Key Laboratory of Forensic Science and Technology at College of Sichuan Province (2023YB04).

Author Contributions: Study conception and design: Song Liu, Hongying Zhang; data collection: Song Liu, Xue Li; analysis and interpretation of results: Song Liu, Xi Yang; draft manuscript preparation: Song Liu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data used in this paper can be requested from the corresponding author upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 593–600, 2007. doi: 10.1109/TCE.2007.381734.

[2] Celik and T. Tjahjadi, "Contextual and variational contrast enhancement," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3431–3441, 2011. doi: 10.1109/TIP.2011.2157513.

[3] H. -D. Cheng and X. Shi, "A simple and effective histogram equalization approach to image enhancement," *Digit. Signal Process.*, vol. 14, no. 2, pp. 158–170, 2004. doi: 10.1016/j.dsp.2003.07.002.

[4] C. Lee, C. Lee, and C. -S. Kim, "Contrast enhancement based on layered difference representation of 2D histograms," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5372–5384, 2013. doi: 10.1109/TIP.2013.2284059.

[5] E. D. Pisano *et al*., "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *J. Digit. Imaging*, vol. 11, no. 4, pp. 193–200, 1998. doi: 10.1007/BF03178082.

[6] Z. -G. Wang, Z. -H. Liang, and C. -L. Liu, "A real-time image processor with combining dynamic contrast ratio enhancement and inverse gamma correction for PDP," *Displays*, vol. 30, no. 3, pp. 133–139, 2009. doi: 10.1016/j.displa.2009.03.006.

[7] S. -C. Huang, F. -C. Cheng, and Y. -S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1032–1041, 2012. doi: 10.1109/TIP.2012.2226047.

[8] 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, 2016. doi: 10.1186/s13640-016-0138-1.

[9] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Optical Society. America*, vol. 61, no. 1, pp. 1–11, 1971. doi: 10.1364/JOSA.61.000001.
[10] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using CNNs," in *BMVC*, Northumbria University, 2018, vol. 220, no. 1.
[11] S. Moran, P. Marza, S. McDonagh, S. Parisot, and G. Slabaugh, "DeepLPF: Deep local parametric filters for image enhancement," in *Proc. IEEE/CVF Conf. Compu. Vis. Pattern Recognit.*, 2020, pp. 12826–12835.
[12] R. Wang, Q. Zhang, C. -W. Fu, X. Shen, W. -S. Zheng and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6849–6857.
[13] S. W. Zamir *et al.*, "Learning enriched features for real image restoration and enhancement," in *Comput. Vis.–ECCV 2020*, Glasgow, UK, Springer, Aug. 23–28, 2020, pp. 492–511.
[14] 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.*, 2023, pp. 12504–12513.
[15] J. Yao, T. Wu, and X. Zhang, "Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with CNN," 2023, *arXiv:2308.08333*.
[16] X. Zhang, Y. Zhao, C. Gu, C. Lu, and S. Zhu, "SpA-Former: An effective and lightweight transformer for image shadow removal," in *2023 Int. Joint Conf. Neural Netw. (IJCNN)*, IEEE, 2023, pp. 1–8.
[17] Z. Zhang, Y. Jiang, J. Jiang, X. Wang, P. Luo and J. Gu, "Star: A structure-aware lightweight transformer for real-time image enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4106–4115.
[18] Q. Zhao, X. Zhang, H. Tang, C. Gu, and S. Zhu, "Enlighten-anything: When segment anything model meets low-light image enhancement," 2023, *arXiv:2306.10286*.
[19] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Trans. Image Process.*, vol. 30, pp. 2072–2086, 2021. doi: 10.1109/TIP.2021.3050850.
[20] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1013–1037, 2021. doi: 10.1007/s11263-020-01407-x.
[21] T. Ma, C. Fu, J. Yang, J. Zhang, and C. Shang, "RF-Net: Unsupervised low-light image enhancement based on Retinex and exposure fusion," *Comput. Mater. Contin.*, vol. 77, no. 1, pp. 1103–1122, 2023. doi: 10.32604/cmc.2023.042416.
[22] C. Guo *et al.*, "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1780–1789.
[23] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10561–10570.
[24] 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.*, 2022, pp. 5637–5646.
[25] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding and K. -K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22252–22261.
[26] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis., Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.
[27] A. M. Reza, "Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement," *J. VLSI Signal Process. Syst. Signal, Image and Video Technol.*, vol. 38, no. 1, pp. 35–44, 2004. doi: 10.1023/B:VLSI.0000028532.53893.82.
[28] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Process.*, vol. 129, no. 12, pp. 82–96, 2016. doi: 10.1016/j.sigpro.2016.05.031.
[29] X. Fu, D. Zeng, Y. Huang, X. -P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recogni.*, 2016, pp. 2782–2790.

[30] 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, 2016. doi: 10.1109/TIP.2016.2639450.
[31] Z. -U. Rahman, D. J. Jobson, and G. A. Woodell, "Retinex processing for automatic image enhancement," *J. Electron. Imaging*, vol. 13, no. 1, pp. 100–110, 2004. doi: 10.1117/1.1636183.
[32] 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, 2013. doi: 10.1109/TIP.2013.2261309.
[33] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, no. 6, pp. 650–662, 2017. doi: 10.1016/j.patcog.2016.06.008.
[34] Y. Fu, Y. Hong, L. Chen, and S. You, "LE-GAN: Unsupervised low-light image enhancement network using attention module and identity invariant loss," *Knowl. Based Syst.*, vol. 240, no. 6, 2022, Art. no. 108010. doi: 10.1016/j.knosys.2021.108010.
[35] A. Sharma and R. T. Tan, "Nighttime visibility enhancement by increasing the dynamic range and suppression of light effects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11977–11986.
[36] Y. Jiang *et al.*, "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021. doi: 10.1109/TIP.2021.3051462.
[37] P. Wang *et al.*, "Understanding convolution for semantic segmentation," in *2018 IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, IEEE, 2018, pp. 1451–1460.
[38] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.
[39] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1632–1640.
[40] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process Syst.*, vol. 30, 2017, pp. 5998–6008.
[41] 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.*, 2022, pp. 17714–17724.
[42] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested u-net architecture for medical image segmentation," in *Deep Lear. Med. Image Anal. Multimodal Learn. Clinical Decis. Support*. Granada, Spain, Springer, Sep. 20, 2018.
[43] 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.*, 2022, pp. 5901–5910.
[44] J. Bai, Y. Yin, and Q. He, "Retinexmamba: Retinex-based mamba for low-light image enhancement," 2024, *arXiv:2405.03349*.
[45] S. Yang, M. Ding, Y. Wu, Z. Li, and J. Zhang, "Implicit neural representation for cooperative low-light image enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 12918–12927.
[46] Z. Liang, C. Li, S. Zhou, R. Feng, and C. C. Loy, "Iterative prompt learning for unsupervised backlit image enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8094–8103.
[47] X. Zhang *et al.*, "Enlighten-your-voice: When multimodal meets zero-shot low-light image enhancement," 2023, *arXiv:2312.10109*.
[48] R. Al Sobbahi and J. Tekli, "Comparing deep learning models for low-light natural scene image enhancement and their impact on object detection and classification: Overview, empirical evaluation, and challenges," *Signal Process.: Image Commun.*, vol. 109, no. 12, 2022, Art. no. 116848. doi: 10.1016/j.image.2022.116848.
[49] K. Ang, W. T. Lim, Y. P. Loh, and S. Ong, "Noise-aware zero-reference low-light image enhancement for object detection," in *2022 Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, IEEE, 2022, pp. 1–4.