PromptHDR: a low-light image enhancement algorithm based on prompt learning

Wenzhen Yan[1], Yingzhen Wang[1], Fuming Qu[1]

[1] University of Science and Technology Beijing, Beijing, China

Corresponding Author:

Fuming Qu[1]

University of Science and Technology Beijing, Beijing, China

Email address: qufuming@ustb.edu.cn

Abstract: Low-light image enhancement technology is designed to improve the problem of high noise and color distortion of images in low-light environments. Traditional methods, such as Retinex theory, can improve the visual effect but tend to lead to detail distortion, while Transformer algorithm is computationally intensive and suffers from loss of image features and degradation of the efficiency of light information transmission as the distance of information transmission becomes longer. To solve these problems, we propose the PromptHDR algorithm. For the problem of detail distortion, our model firstly uses the Retinex theory to provide the lighting information of the image, and then designs the "U-shape" Transformer architecture to optimize the visual effect of the image through efficient feature extraction and image reconstruction to better capture the detail information, and thus improve the quality of the image. In order to improve the computational efficiency and enhance the generated features, we design a Prompt block for the Transformer module, adopt the Mamba algorithm to improve the efficiency of image processing, and design a powerful sampling module for image detail processing. Quantitative and qualitative experiments on the LOL dataset show that our algorithm outperforms most low-light image enhancement algorithms, proving its effectiveness in low-light image enhancement. Code is now available at: https://github.com/darkforest1908/PromptHDR

Keywords: Low-light Image Enhancement; Sampling; Retinex theory; Transformer algorithm

1 Introduction

Low-light image enhancement is a challenging task in computer vision. Images captured in low light conditions are often accompanied by high noise, low contrast, and color distortion, which significantly reduce the usability of the images[1]. The main goal of low-light image enhancement is to maintain the naturalness of details such as background and color while improving image visibility and visual quality.

Traditional methods for low light enhancement include histogram equalization[2, 3], gamma correction[4], Retinex theory, and Multi-Scale Retinex[5] . Histogram equalization and gamma correction enhance the contrast by adjusting the pixel values of the image, thus improving the visual quality of the image. The Retinex theory, on the other hand, is based on the theory of color constancy in the human visual system and decomposes the image into illumination and reflection components, which can deal with illumination disparity and color distortion at the same time. Multi-Scale Retinex is an extension of basic Retinex theory that applies the Retinex theory at different scales and then fuses the results to better balance local and global enhancements. Although these methods are effective in specific scenes, they generally suffer from insufficient adaptive capability, the need for manual parameterization, difficulty in handling complex low-light images, and difficulty in balancing detail enhancement and naturalness.

Deep learning-based low-light image enhancement algorithms have made significant progress in recent years, and these methods utilize the powerful learning capability of neural networks to automatically extract image features and perform enhancement, which can be mainly classified into two categories: supervised learning and unsupervised learning. Supervised learning methods require pairs of low-light and normal-light images as training data, and learn the enhancement mapping by minimizing the difference between the prediction result and the target image. Representative supervised learning algorithms include CNN-based LLNet[6] and MBLLEN[7], as well as RetinexNet[8] and KinD[9], which incorporate Retinex theory. These methods usually yield high-quality enhancement results, but rely on a large amount of labeled data and may have limited generalization ability in unseen scenarios.

In contrast, unsupervised learning methods do not require paired training data, but instead implement self-supervised learning through a specific loss function or network structure. Typical unsupervised methods include GAN-based EnlightenGAN [10], RDGAN [11], and zero-reference-based Zero-DCE[12]. These methods overcome the data dependency problem and usually have better generalization ability, but the overall enhancement effect is not as stable as supervised methods due to the lack of scenario-specific training.

In recent years, Transformer-based methods have shown excellent results in both supervised and unsupervised domains. Algorithms such as STAR [13]and EndoUIC [14]utilize the self-attention mechanism of the Transformer algorithm to capture the long-range dependencies of the image, and perform well in dealing with the global illumination disparity problem. These methods bring new ideas for low-light image enhancement to better handle complex lighting conditions, but also face the challenge of high computational complexity.

To address the shortcomings of existing algorithms, we propose PromptHDR——a Transformer-based low-light image enhancement algorithm. The algorithm is divided into two main stages:

1. In the first phase, the illumination extraction module was designed to address the problem of color distortion after processing images by traditional algorithms. The low-light images are first processed to extract the lighting features and mapped back to the original image space to generate the final illumination mapped image.

2. In the second phase, The Transformer architecture has the problems of redundant long-distance information transfer, inefficient information processing, and lack of direct modeling of local spatial features, which may affect the image details, etc. The “U-shaped” connection structure is designed based on the Transformer architecture. Based on the Transformer architecture, a “U-shape” connection structure is designed to make full use of the contextual information, and the ability of the model to extract features is improved by introducing the Prompt block. In addition, a powerful sampling module is designed to solve the problem of local detail loss caused by global feature modeling and improve the accuracy of image details. Finally, the illumination mapping image generated in the first stage is fused with the error components output in the second stage to generate the final enhanced image.

To comprehensively evaluate the performance of PromptHDR, we conducted qualitative and quantitative tests on the LOL low-light dataset. The results show that PromptHDR outperforms most of the current deep learning algorithms based on the LOL dataset on this dataset to a leading level.

Our main contributions can be summarized as follows:

1. A light extraction module based on Retinex physical model is introduced to extract the light feature information of low-light images in the preprocessing stage, which provides richer light data for subsequent processing.

2. The Prompt block is designed. It consists of two parts: The Global Context Prompting part captures the overall scene information and generates guidance cues; and The Hierarchical Feature Refinement part fuses the cue information and lighting features, balances the detailed and

global information, and optimizes the image features.

3. A powerful sampling module has been introduced to improve the performance of the Transformer module in extracting features. Downsampling extracts multi-scale information, preserves global structure and reduces noise, while upsampling balances global and local features and improves overall recovery details.

2 Related Work

2.1 Low-light image enhancement

The development of low-light image enhancement technology in recent years lies mainly in two key stages: deep learning combined with Retinex and Transformer model application. In the deep learning combined with Retinex stage, researchers have combined Retinex theory with deep neural networks to create a series of innovative methods. Among them, RetinexNet deals with the problem of uneven illumination and color distortion through CNN, while GAN-based methods such as also show excellent performance. With the deepening of the research, the introduction of the Transformer model opens a new stage of development, providing stronger global dependency modeling capabilities. LLFormer [15]utilizes the self-attention mechanism to capture long-range dependencies, RetinexFormer [16] combines the Retinex theory with the Transformer model to achieve excellent enhancement results, and IAT [17], Restormer [18], and MIRNet [19]have further explored the Transformer in low-light image enhancement, such as adding adaptive attention mechanism and combining with U-Net structure, have achieved remarkable results.

2.2 Prompt Learning

The learning prompt network originated in the field of natural language processing, which improves learning efficiency by providing contextual information for the model. Subsequently, this concept was successfully extended to visual tasks and incremental learning, showing significant optimization effects. The learning prompt network is mainly divided into two categories: insertion prompt module and generation prompt module. Insertion modules, such as Pro-Tuning [20]and MIRNet, mainly add prompt processing between network layers. Generative modules, such as CSVPT [21], generate feature prompt sequences through independent networks. These methods are all designed to enhance the performance of the model by providing additional contextual information. In addition, CoOp [22] proposed a visual language cue learning method based on context optimization, which converts the context words in the cue into a set of learnable

vectors for learning. HiPro [23] makes the pre-trained visual language model adapt to other downstream tasks by constructing a hierarchical task tree and combining sharing and personalized tips. These methods provide new ideas for the research in the field of image processing, so that the model can better meet the needs of different image tasks, while showing higher efficiency in fine-tuning and incremental learning.

2.3 State space models

Mamba [24]is an emerging deep learning architecture based on the idea of Structured State Space Models (SSMs) [25], which aims to solve the computational bottleneck of Transformer when processing long sequences. By utilizing the linear time complexity of SSM, Mamba can improve computational efficiency while maintaining the expressive power of the model as much as possible, thus processing long sequence tasks more efficiently. Recent studies have shown that Mamba shows significant potential in the field of image processing, especially in tasks such as image classification, image restoration, and biological and medical image segmentation. Based on the excellent performance of Mamba, some scholars have further expanded its application. For example, Vmamba [26]designed a Cross-Scan module to enable the model to selectively scan 2D images, which greatly improves the efficiency and speed of image processing. In addition, Mamba 's research results have been widely used in other fields, such as biological research [27]and traffic analysis [28], demonstrating its strong adaptability and potential for innovation in multiple disciplines.

3 Method

The overall architecture of our proposed PromptHDR is shown in *Fig. 1* and consists of two parts, Lighting Extraction Module (LEM) and Feature Fusion Module (FFM).

The input image is first passed through a light estimator to generate illumination maps and features, which are then fed into the FFM for further extraction. The FFM adopts a “U-shape” network structure consisting of multiple layers of transformers, which enhances the image details and luminance performance through the joint extraction of global and local features.

The first three layers of Transformer combine the downsampling module to retain the global semantics while extracting multi-scale local features. Subsequently, the three-layer Transformer is connected to the Prompt module and the upsampling module; the Prompt block guides, fuses and refines features to further optimize features, improve the quality of low-light images, and enhance details and naturalness. The upsampling module recovers local details and splices the

upsampled features with the transformer output to enhance the contextual information acquisition capability. In between, the features of different layers are also spliced through the “U-shaped” connection structure to improve the network's ability to process contextual information and enhance its feature extraction capability.

Finally, the depth feature is extracted by the two-layer Transformer, and then processed by the out converter, and stitched with the initial input image to output the enhanced image.

Fig. 1

Fig. 1 The overall architecture of the proposed PromptHDR is as follows: (a) This method comprises two primary stages. The first stage is the Light Extraction Module (LEM) (i), which extracts illumination priors from the input image and combines them with the original RGB image. Lightweight convolutional operations, including pointwise convolution and depthwise separable convolution, are employed to extract spatial features, resulting in the generation of an illumination map and intermediate illumination features. These outputs serve as inputs to the subsequent Prompt module for feature optimization. The second stage is the Feature Fusion Module (FFM) (ii), which adopts an improved U-shaped network structure. It utilizes a Transformer to capture both global semantic information and local details of the image, incorporating multi-scale downsampling, prompt modules, and upsampling mechanisms to achieve efficient feature fusion and enhancement. Ultimately, the fused deep features are concatenated with the initial input image, producing an enhanced image with superior quality. (b) The structure of the Transformer Block is illustrated on the right side of *Fig. 2*, while the Prompt module consists of the Global Context Prompt (GCP) and High-Frequency Prompt (HFP), designed to improve global illumination perception and local detail extraction, as depicted at the bottom of *Fig. 2* (c) The sampling module includes downsampling (HWD) and upsampling (DySample), which facilitate multi-scale feature extraction and reconstruction. Further details are provided in Section 3.3.

3.1 Lighting Extraction Module

The specific structure of the illumination extraction module is shown in *Fig. 1* (I) : First, the network receives an RGB image as input. The first step is to calculate the illumination prior, which is achieved by averaging all the color channels of the input image. Next, the original RGB image and the calculated illumination prior are stitched together, and then the input data is reconstructed between channels through the first 1x1 convolution layer. The resulting feature map is passed through a 5x5 deep separable convolutional layer, which allows the network to

efficiently extract spatial features and capture local lighting patterns and changes. Finally, the feature map is used to map the extracted features back to the original image space through the second 1x1 convolution layer to generate the final illumination mapping. The final result output illumination map is used for feature extraction of the lower-level network, and the middle illumination feature is used to optimize the features of the Prompt block. The specific formula is expressed as:

Given an input RGB image I : (b, 3, h, w) ， Calculating Lighting Prior Lp：

$$\mathrm{Lp} = \left(\frac{1}{3}\right) * \sum(\mathrm{R}, \mathrm{G}, \mathrm{B}) \tag{1}$$

The result is output as feature X：

$$\mathrm{X} = \mathrm{f(I)} = \mathrm{Conv2}\Big(\mathrm{DepthConv}\big(\mathrm{Conv1}(\mathrm{I} \oplus \mathrm{Lp})\big)\Big) \tag{2}$$

Where $\oplus$ Indicates a splice operation; $Conv1$、$DepthConv$ and $Conv2$ represents three convolution operations respectively ;

3.2 Prompt block based on Mamba and lighting features

The structure of the Prompt block is shown in the lower part of *Fig. 1*: the module consists of two phases, Global Context Prompting and Hierarchical Feature Refinement. The first phase focuses on the capture of global information and the generation of prompts, and enhances the model's understanding of the overall structure of the scene through the generation of guided prompting information. In the second stage, through multi-level feature fusion and self-attention refinement processing, the cue information and lighting features are highly fused, and a balance is achieved between the details and global information, and finally enhanced features are generated for image enhancement. The formulas for the two stages are represented as follows:

$$\mathrm{O}\big(\mathrm{x}_{in}, \mathrm{I(x)}\big) = \mathrm{HFR}\left(\mathrm{Cat}\big(\mathrm{x}_{in}, \mathrm{P}(\mathrm{x}_{in})\big), \mathrm{I}(\mathrm{x}_{in})\right) \tag{3}$$

Where $O(\mathrm{x}_{in}, I(x))$ is the final output of the entire Prompt block.

x_{in} is the input feature, and $I(x)$ is the light feature generated in the first stage, and $P(\mathrm{x}_{in})$ is the cue message generated by the Prompt block. $Cat(\mathrm{x}_{in}, P(\mathrm{x}_{in}))$ denotes the splicing of the input features with the prompt information in the channel dimension. Then after the HFR module fuses the light features $I(x)$ and splicing, after further refinement, the final output features are obtained.

Phase I GCP Process: Where $P(x_{in})$ denotes the first stage of GCP processing, which

extracts global information from input features, generates cued features by cueing parameters and weights, followed by resizing and convolutional processing, and outputs guided cueing information. This stage is mainly used to capture the global context and enhance the effectiveness of subsequent feature processing. The formula is expressed as:

$$P(x_{in}) = Conv3x3\left(\mathrm{Interp}\left(\sum_{i=0}^{L} \omega_i \cdot P_i\right)\right) \tag{4}$$

Where P_i denotes the i cue generated from the parameter P ; ω_i is the weight calculated by the linear layer with Softmax, indicating the importance of each cue; $Interp(\cdot)$ Indicates that the cue is scaled to the size of the input feature map by bilinear interpolation;

Phase II HFR Process: Fuse the Prompt block output first with the light feature:

$$I = \mathrm{Conv1x1}(I(x)) \tag{5}$$

$$P'(x_{in}) = \mathrm{Norm}(P(x_{in})) + \mathrm{Attention}(\mathrm{Norm}(P(x_{in})), I) \tag{6}$$

where $I(x)$ is the illumination feature and $P(x)$ is the input feature.

Then, the fusion features are enhanced by the feedforward network (FFN) :

$$P''(x_{in}) = P'(x_{in}) + FFN\left(Norm(P'(x_{in}))\right) \tag{7}$$

Final refinement of features and upsampling to recover spatial information.

$$x_{out} = P''(x_{in}) + DropPath\left(SS2D\left(Norm(P''(x_{in}))\right)\right) \tag{8}$$

Where $SS2D$ is the self-attention mechanism, and $DropPath$ used to introduce random depth.

3.3 Sampling

The performance of the “U-shaped” Transformer structure in low-light image enhancement tasks can be significantly improved by the DySample [29]and HWD [30]modules that process the features at the appropriate stage. The HWD module extracts multi-scale information through wavelet transform, effectively preserving the global structure and detailed features, which makes up for Transformer's lack of local detail processing. The low-frequency components after downsampling help Transformer to focus on global features and reduce noise interference, while the DySample module accurately recovers the details by dynamically adjusting the sampling position and enhances the detail performance in blurred or dark areas. In addition, this combination of up and downsampling achieves a balance between global and local features, and effectively controls the computational overhead to improve the efficiency of low-light data

processing.

3.3.1 UpSampling

Fig. 2

Fig. 2 Structure of DySample

The DySample module implements a dynamic sampling mechanism, the core idea of which is to flexibly adjust the spatial sampling position by learning the offsets of the input features. Specifically, the module generates offsets through two parallel convolution operations: one for generating the original offsets, the offset convolution. The other is an optional range convolution that is used to modulate these offsets to generate the final sampling positions.

First, the input feature maps are convolved with offsets to generate raw offsets. These raw offsets are further processed to obtain the final offsets. The optional range convolution then modulates the offsets to generate the initial sampling locations.

Next, the offsets are combined with the standard pixel grid to generate the final sampling coordinates. The sampling coordinates are calculated by the following formula:

$$coords = \frac{2 * \left(grid + processed_{offset}\right)}{[W, H]} - 1 \tag{9}$$

Where $grid$ is a standard pixel grid, divided by $[W, H]$ for normalization, and then the range is adjusted to [-1,1].

Finally, these coordinates are used to sample the input:

$$y = F.grid_{sample} \tag{10}$$

This operation uses bilinear interpolation to sample at the given coordinates.

The following is the overall formula of Dysample:

$$y = DySample(x) = F.grid_{sample(x,coords)} \tag{11}$$

This process dynamically adjusts the spatial sampling positions by learning the offsets of the input features, thus enhancing the network's ability to adapt to spatial information transformations.

3.3.2 Downsampling

Fig. 3

Fig. 3 Structure of HWD module

The HWD module further preserves and enhances the multi-scale local feature representation through downsampling after combining the global semantic information extracted

by Transformer. The workflow of the module can be divided into the following two main steps:

1. Wavelet Decomposition and Feature Splicing:

First, the input features are decomposed into one low-frequency component and three high-frequency components corresponding to horizontal, vertical, and diagonal detail information by the Discrete Wavelet Transform (DWT). The low-frequency component contains the global structural information of the input, while the high-frequency components capture more detailed features. Splicing these components to form a multi-scale feature tensor can provide richer semantic information for subsequent feature transformations. The process can be represented by the following equation:

$$x_{cat} = Concat([y_L, y_{HL}, y_{LH}, y_{HH}]) \tag{12}$$

Where y_L is the low frequency component, and y_{HL}, y_{LH} and y_{HH} are the high frequency components in the horizontal, vertical and diagonal directions, respectively.

2. Feature transformation and expression enhancement:

The spliced feature tensor undergoes a series of operations, including point-by-point convolution (equivalent to 1x1 convolution) for adjusting the number of channels, batch normalization for optimizing the feature distribution and improving the stability of the model, and ReLU nonlinear activation function for enhancing the feature representation. The final downsampled output is represented by the following equation:

$$y = DWT(x) = ReLU\left(BatchNorm\left(Conv1x1(x_{cat})\right)\right) \tag{13}$$

This process not only preserves the low and high frequency information, but also enhances the expressive ability of the feature map through the transform.

Through this series of processing, the HWD module is able to retain the high-frequency and low-frequency information in the input data while using the wavelet transform to strengthen the modeling ability of multi-scale features, and ultimately generates downsampled features with both global semantics and local details. This provides richer feature representations for subsequent tasks and improves the performance of the model.

4 Experiments

4.1 Data sets and implementation details

Fig. 4

Fig. 4 Comparison of 3D effects in HSV color space of enhanced images

We train and test the proposed PromptHDR algorithm on the LOL dataset.

4.1.1 Data sets

We use the LOL dataset, which is widely used in low-light image enhancement research, to train and test the model. the LOL dataset is divided into two versions, v1 and v2, covering diverse low-light conditions in both real and synthetic scenes. LOL-v1 contains 485 pairs of high and low light images for training and 15 pairs of images for testing; LOL-v2 is further extended into two subsets: real and synthetic. real subset is captured in a real scene by adjusting the ISO and the exposure time, and contains 689 pairs of images reflecting the lighting variations in the actual scene; synthetic subset, on the other hand, simulates complex low-light conditions through specific synthesis techniques and contains 1000 pairs of images, of which 900 pairs are used for training and 100 pairs are used for testing. These data provide a comprehensive validation basis for model adaptation and enhancement in diverse scenes.

4.1.2 Implementation Details

The computer configuration for this experiment is Windows 11 (CUDA11.8, Python3.10, Pytorch11.2), GPU is NVIDIA RTX A6000, CPU is Intel i9-13900K processor. The resolution of our images was set to 128 × 128 and optimized using the Adam optimizer, where β_1 was set to 0.9, the RMSprop control parameter β_2 was set to 0.999 and cosine annealing was used to prevent falling into local minima for a total of 150,000 iterations.

4.2 Low-light Image Enhancement

In *Table 1*, we present a comparison of the results of our method with those of the supervised and unsupervised methods for low-light image enhancement. The test images are derived from the LOL dataset. The data in the table are trained and tested using the authors' public code or obtained from the original paper under the same configuration parameters. The experimental results show that our algorithm outperforms other algorithms for SOTA in the test evaluation metrics such as PSNR, SSIM, and LPIPS.

Table 1 Performance of each algorithm on the LOL dataset

Table 1

4.2.1 Quantitative results

The metrics we use to evaluate the algorithms are SSIM, PSNR, and LPIPS. The following is a description of each evaluation metric:

1. SSIM（Structural Similarity Index Measure） is a measure of the similarity between two images that takes into account the structural information of the image, which makes it more in line with the way the human eye perceives it when evaluating the quality of an image. The value of SSIM ranges from -1 to 1, where 1 means that the two images are exactly the same, and 0 means that they are completely different. The closer the SSIM is to 1, the better the quality of the image or the more structural similarity between the two images. The formula is as follows:

$$SSIM(X,Y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{14}$$

Where μ_x and μ_y are the mean values of image x and image y; σ_x^2 and σ_y^2 are the variance of image x and image y; σ_{xy} is the covariance of image x and image y; C_1 and C_2 are constants used in stabilization calculations and are usually taken to be small to avoid a zero denominator.

2. PSNR（Peak Signal-to-Noise Ratio）is a commonly used image quality evaluation metric to measure the quality of image compression or reconstruction. It is mainly used to compare the difference between the original image and the image after processing (e.g. compression or denoising.) The higher the PSNR value, the better the image quality and the lower the distortion. The formula is as follows:

$$MSE = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}[I(i,j) - K(i,j)]^2 \tag{15}$$

Where $I(i,j)$ and $K(i,j)$ the pixel values at position (i,j) for the original and processed images respectively. R is the dynamic range of the image pixel value, usually 255.

$$PSRN = 10 \cdot \log_{10}(\frac{R^2}{MSE}) \tag{16}$$

3. LPIPS（Learned Perceptual Image Patch Similarity）is a perception-based image quality assessment metric that evaluates the similarity between images by learning the perceptual distances of image features to provide a quality metric that is more in line with human visual perception. The formula is expressed as follows:

Features are first extracted, for the original image I and the processed image K. The feature representations $F_i(I)$ and $F_i(K)$ are extracted by the pre-trained network, where i denotes the

feature layer.

The feature distances are then computed and weighted for summation. The feature distances of each layer are weighted using the weights ω_i and summed to get the final LPIPS value:

$$\mathrm{LPIPS}(\mathrm{I},\mathrm{K}) = \sum_{\mathrm{i}} \omega_{\mathrm{i}} \cdot \| F_i(I) - F_i(K) \|_2 \tag{17}$$

Where: $\|\cdot\|_2$ denotes the Euclidean distance (L2 paradigm); The weights ω_i are obtained by learning from training data subjectively evaluated by humans.

A comparison of the evaluation metrics of each algorithm is shown in *Table 1*, and the results show that our proposed algorithm outperforms most of the low-light enhancement algorithms and achieves the best enhancement results.

4.2.2 Qualitative results

Fig. 5

Fig. 5 Comparison of the effect of each algorithm in the LOL-v1 dataset, our method effectively reduces noise and color distortion

Fig. 6

Fig. 6 Comparison of the effect of each algorithm in LOL-v2-real dataset, our method effectively reduces noise and color distortion

The qualitative results of our PromptHDR algorithm compared with other SOTA algorithms are shown in *Fig. 4* and *Fig. 5* above. On the LOLv1 dataset, the color restoration degree and edge sharpening degree of our algorithm have achieved the best results. Compared with EnlightenGan and KinD, the color is dark and the noise of Zero-DCE is large, and Uretinex-Net has overexposure. Similarly, in *Fig. 5*, the PromptIR digital color restoration is insufficient, while the Uretinex-Net and KinD digital noise are large. On the whole, the enhancement effect of our algorithm on LOLv1 and LOLv2-real datasets has achieved excellent results in color restoration, noise control and color distortion reduction.

We also compare and analyze the normal illumination, low illumination and enhanced images in the same scene based on HSV color space. The results are shown in *Fig. 4* and *Fig. 7*. The distribution of the enhanced image is similar to that of the normal illumination image, and the V channel is significantly improved. At the same time, the structural consistency of the H and S distribution is maintained, which verifies the brightness recovery and color restoration effect of the enhancement method.

Fig. 7

Fig. 7 Comparison of 2D effect of HSV color space of enhanced image

4.3 Ablation Study

Table 2 Results of the ablation experiments

Table 2

We set up eight architectures to perform ablation experiments on our own code by removing the relevant components for validation on the LOL-v1 dataset, and the results of the experiments are shown in *Table 2* Compared with the other architectures of the ablation experiments, our final combined architecture obtained the best performance.

Fig 8 shows the performance comparison of different architectures on the LOL-v1 dataset. The visual comparison shows that architecture 3, which is our proposed PromptHDR algorithm, generates images with sharper brightness and contrast, significant detail recovery, good performance in shadow and highlight areas, avoiding overdarkness or overexposure, and the overall image is natural and visually comfortable.

Fig. 8

Fig. 8 Comparison effect of ablation experiment pictures, Architecture 1: IE represents the architecture containing only Illumination Estimator, Architecture 2: IE+P represents the architecture containing both Illumination Estimator and Prompt, Architecture 3: IE+P+S represents the architecture containing Illumination Estimator, Prompt and Sampling.

In order to verify the validity of the choice of internal constructs for the sampling module, we also performed ablation experiments on it, separately verifying the effectiveness of the upsampling module and the downsampling module on the LOL-v1 dataset. The results are shown in *Table 3*, and the training results of both plus are significantly better than plus sampling or downsampling alone.

Table 3 Sampling module ablation experiments

Table 3

5 Conclusion

In this paper, we present the PromptHDR architecture based on Retinex theory and Transformer block for low-light image enhancement. The study is based on the “U-shaped” Transformer architecture, which introduces an illumination extraction block to obtain more illumination information for the input image. The Transformer block with Prompt hints further improves the accuracy of the feature representation, while the sampling module effectively

enhances the image details. Extensive experiments show that our algorithm outperforms most existing methods on the LOL dataset.

Although the model performs well in terms of overall enhancement effect, its complex network architecture is the main challenge. Future work will focus on further improving the accuracy of the model while optimizing the network architecture and reducing the number of parameters for more efficient image enhancement.

References

1. Liu, X., et al., *NTIRE 2024 Challenge on Low Light Image Enhancement: Methods and Results.* ArXiv, 2024. abs/2404.14248.
2. Pizer, S.M., et al., *Adaptive histogram equalization and its variations.* Computer Vision, Graphics, and Image Processing, 1987. 39(3): p. 355-368.
3. Pisano, E.D., et al., *Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms.* Journal of Digital imaging, 1998. 11: p. 193-200.
4. Rahman, S., et al., *An adaptive gamma correction for image enhancement.* EURASIP Journal on Image and Video Processing, 2016. 2016(1): p. 35.
5. Petro, A.B., C. Sbert, and J.-M. Morel, *Multiscale retinex.* Image processing on line, 2014: p. 71-88.
6. Lore, K.G., A. Akintayo, and S. Sarkar, *LLNet: A deep autoencoder approach to natural low-light image enhancement.* Pattern Recognition, 2017. 61: p. 650-662.
7. Lv, F., et al. *MBLLEN: Low-Light Image/Video Enhancement Using CNNs*. in *British Machine Vision Conference*. 2018.
8. Wei, C., et al., *Deep retinex decomposition for low-light enhancement.* arXiv preprint arXiv:1808.04560, 2018.
9. Zhang, Y., J. Zhang, and X. Guo. *Kindling the darkness: A practical low-light image enhancer*. in *Proceedings of the 27th ACM international conference on multimedia*. 2019.
10. Jiang, Y., et al., *Enlightengan: Deep light enhancement without paired supervision.* IEEE transactions on image processing, 2021. 30: p. 2340-2349.
11. Wang, J., et al. *RDGAN: Retinex decomposition based adversarial learning for low-light enhancement*. in *2019 IEEE international conference on multimedia and expo (ICME)*. 2019. IEEE.

12. Guo, C., et al. *Zero-reference deep curve estimation for low-light image enhancement*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
13. Zhang, Z., et al. *Star: A structure-aware lightweight transformer for real-time image enhancement*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
14. Bai, L., et al., *EndoUIC: Promptable Diffusion Transformer for Unified Illumination Correction in Capsule Endoscopy.* arXiv preprint arXiv:2406.13705, 2024.
15. Wang, T., et al. *Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023.
16. Cai, Y., et al. *Retinexformer: One-stage retinex-based transformer for low-light image enhancement*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
17. Cui, Z., et al., *You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction.* arXiv preprint arXiv:2205.14871, 2022.
18. Zamir, S.W., et al. *Restormer: Efficient transformer for high-resolution image restoration*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
19. Zamir, S.W., et al., *Learning enriched features for fast image restoration and enhancement.* IEEE transactions on pattern analysis and machine intelligence, 2022. 45(2): p. 1934-1948.
20. Nie, X., et al., *Pro-tuning: Unified prompt tuning for vision tasks.* IEEE Transactions on Circuits and Systems for Video Technology, 2023.
21. Li, A., et al. *Learning common and specific visual prompts for domain generalization*. in *Proceedings of the Asian Conference on Computer Vision*. 2022.
22. Zhou, K., et al., *Learning to Prompt for Vision-Language Models.* International Journal of Computer Vision, 2022. 130(9): p. 2337-2348.
23. Liu, Y., et al. *Hierarchical prompt learning for multi-task learning*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
24. Gu, A. and T. Dao, *Mamba: Linear-time sequence modeling with selective state spaces.* arXiv preprint arXiv:2312.00752, 2023.
25. Gu, A., et al., *Combining recurrent, convolutional, and continuous-time models with linear state space layers.* Advances in neural information processing systems, 2021. 34: p.

572-585.
26. Zhu, L., et al., *Vision mamba: Efficient visual representation learning with bidirectional state space model.* arXiv preprint arXiv:2401.09417, 2024.
27. Shi, R., et al., *ShapeMamba-EM: Fine-Tuning Foundation Model with Local Shape Descriptors and Mamba Blocks for 3D EM Image Segmentation.* arXiv preprint arXiv:2408.14114, 2024.
28. Yuan, D., et al., *ST-Mamba: Spatial-Temporal Mamba for Traffic Flow Estimation Recovery using Limited Data.* arXiv preprint arXiv:2407.08558, 2024.
29. Liu, W., et al. *Learning to upsample by learning to sample*. in *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.
30. Xu, G., et al., *Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation.* Pattern Recognition, 2023. 143: p. 109819.
31. Ma, L., et al. *Toward fast, flexible, and robust low-light image enhancement*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
32. Yang, W., et al., *Sparse gradient regularized deep retinex network for robust low-light image enhancement.* IEEE Transactions on Image Processing, 2021. 30: p. 2072-2086.
33. Liu, R., et al. *Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021.
34. Yang, S., et al. *Implicit neural representation for cooperative low-light image enhancement*. in *Proceedings of the IEEE/CVF international conference on computer vision*. 2023.
35. Yang, W., et al. *From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.
36. Wu, W., et al. *Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
37. Potlapalli, V., et al., *Promptir: Prompting for all-in-one image restoration.* Advances in Neural Information Processing Systems, 2024. 36.