



Row–Column Separated Attention Based Low-Light Image/Video Enhancement

Chengqi Dong,¹ Zhiyuan Cao,² Tuoshi Qi,¹ Kexin Wu,¹ Yixing Gao^{1,*} and Fan Tang³

¹School of Artificial Intelligence, Jilin University, Changchun, China
{dongcq9920, qits9920, wukx21}@mails.jlu.edu.cn, gaoyixing@jlu.edu.cn

²College of Software, Jilin University, Changchun, China
caozy9920@mails.jlu.edu.cn

³Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
tfan.108@gmail.com

Abstract

U-Net structure is widely used for low-light image/video enhancement. The enhanced images result in areas with large local noise and loss of more details without proper guidance for global information. Attention mechanisms can better focus on and use global information. However, attention to images could significantly increase the number of parameters and computations. We propose a Row–Column Separated Attention module (RCSA) inserted after an improved U-Net. The RCSA module's input is the mean and maximum of the row and column of the feature map, which utilizes global information to guide local information with fewer parameters. We propose two temporal loss functions to apply the method to low-light video enhancement and maintain temporal consistency. Extensive experiments on the LOL, MIT Adobe FiveK image, and SDSD video datasets demonstrate the effectiveness of our approach.

Keywords: Low-light image/video enhancement, Row-column separated attention, Temporal consistency

CCS Concepts: • Computing methodologies → Image processing; Computational photography; Computer vision representations

1. Introduction

Limited by realistic objective conditions, we encounter many low-light images and videos. A low-light environment leads to imaging degradations, including intense noise, low visibility, chromatic aberrations, and information loss [XWFJ22a, LLWL18, KYK*17]. Low-light image enhancement is proposed to improve the quality of images captured under low-light conditions. The main goal of the low-light image enhancement algorithm is to create visually pleasing images and provide more information than input images suitable for applications [TXB19, WZHL13]. Low-light data enhancement methods can improve the aesthetic quality of low-light visual media and contribute to data pre-processing of visual tasks [LLWL18, ZLYF21]. The goal of image enhancement is to remove image noise, increase brightness, improve the structural similarity of images, improve visual perception and reduce colour difference.

With the development of deep learning, neural network-based methods [LGH*21, LLWL18, ZAK*20, TTZ*22, FLL22] have emerged. Through a well-designed network, many encoder and decoder models represented by the U-Net structure [ZLYF21, ZLL22, FLL22] can achieve better enhancement results. However, since these methods perform multiple down-sampling operations and only use convolution kernels to extract features, the enhanced results may lose many local details and lack global guidance. This means the enhancement effect worsens in low-light areas with high local noise. Currently, there are also global information guidance methods based on the transformer. Still, the transformer-based methods only accept relatively small patches of fixed sizes, which inevitably cause patch boundary artefacts when applied on larger images using cropping [TTZ*22]. Therefore, the size of the input image is limited, and the number of parameters is large, which means training and inference are more time-consuming. To address the fact that the computational complexity of ViT is quadratic with the resolution of input, LLFormer [WZS*23] proposed A-MSA to compute self-attention

*Corresponding author: gaoyixing@jlu.edu.cn

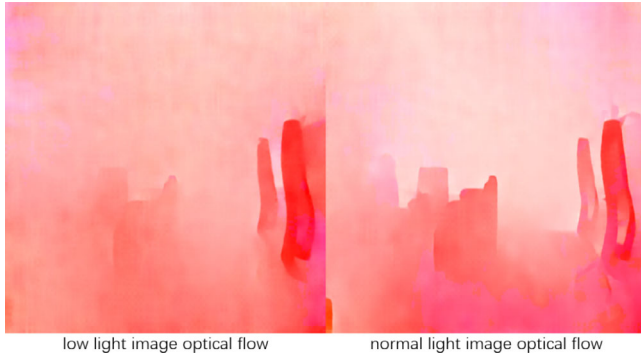


Figure 1: The light flow results of the low and normal light scenes are calculated by the GMflow [XZC*22] model, in which the light flow of the dark scene is more blurred and missing many details.

on the height and width axes across the channel dimension sequentially and employed depth-wise convolutions. STAR [ZJJ*21] flattened the full-size image into a sequence of image blocks and then reduced the spatial resolution by Adaptive Average Pooling to make dimensionality reduction of the image blocks, which significantly reduced the complexity and computation of generating tokens. Other researchers have also conducted research on row and column attention, such as the RCANet [LHWH22] paper used for semantic segmentation, which separately sets row and column attention operations on each channel and then merges them without cross-channel information exchange. The Lanformer [HDC*22] paper used for lane detection uses the row and column feature vectors extracted by the network for attention operations and adds the row and column results as the output of the attention module, which cannot obtain pixel-level attention results. We propose row-column separated attention (RCSA) to extract global information with fewer parameters and faster training. Our RCSA module allows direct cross-channel information interaction, obtains pixel-level attention results and achieves self-balancing between maximum and average attention.

Furthermore, we attempt to apply the method to video enhancement. Nevertheless, directly using image enhancement methods on videos will lead to obvious flicker and jitter [XWFJ22a]. Therefore, the main challenge from image to video is achieving temporal stability of video frames while ensuring the quality of single-frame enhancement. A common approach is to extend 2D image models to 3D video models, such as MBLLEN [LLWL18]. However, replacing the convolution kernel lacks global information and requires much computation. In addition to the 3D convolution model, some methods use optical flow information to enhance the video. They use optical flow to describe the motion of objects in dynamic scenes, such as StableLLVE [ZLYF21]. However, such methods are affected by the accuracy of optical flow estimation. Specifically, calculating optical flow in low-light environments is more challenging, as shown in Figure 1. To ensure the stability of enhancement results, we propose two loss functions based on the characteristics of adjacent video frames. These functions consider the local differences and self-similarity of adjacent frames, respectively.

In summary, considering that the existing methods lack global information guidance and are limited by the transformer methods, we innovatively propose a RCSA module inserted after the improved U-Net. The RCSA module and improved U-Net form the U-RCSA block. Furthermore, we design the U-RCSANet by stacking three U-RCSA blocks with the same parameters as shown in Figure 2. The improved U-Net is responsible for extracting local information and fusing shallow information with deep information in the framework. At the same time, the RCSA module has a small number of parameters and is not limited by image size, which is responsible for conducting guidance on global information. Our contributions are as follows.

- We propose a novel U-RCSANet that leverages global context to guide local information while fusing shallow and deep features for low-light enhancement tasks.
- We propose an efficient and compact parameter RCSA module, which guides pixel-level attention results by learning global feature information of rows and columns.

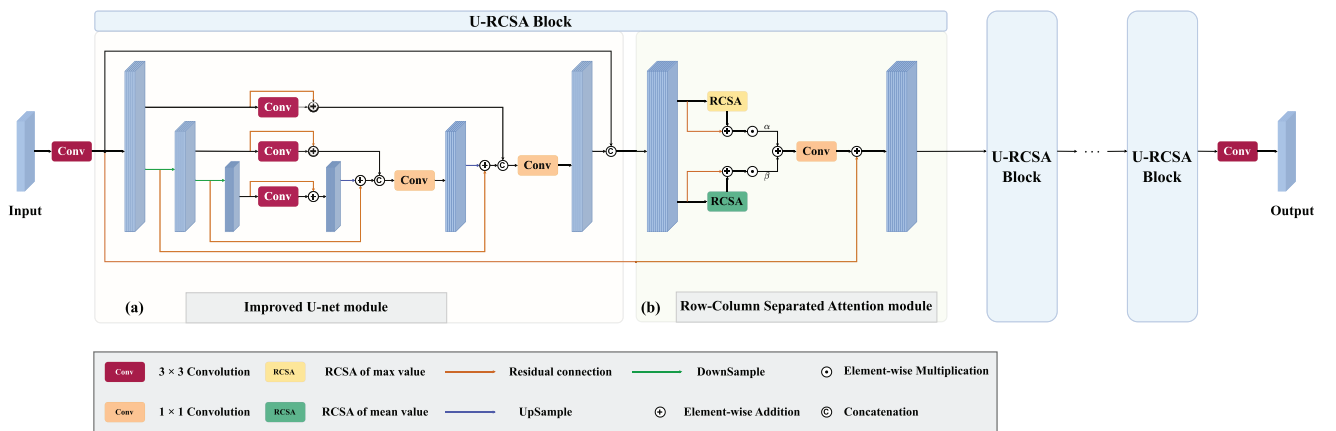


Figure 2: The framework of U-RCSANet. U-RCSANet consists of three U-RCSA blocks with the same parameters. Each U-RCSA block has an improved U-Net and a Row-Column Separated Attention module.

- We propose two temporal loss functions to guide the model in learning the temporal stability information of the video. Extensive experiments on image and video datasets have proved that our method is more effective than the existing models.

2. Related Work

Low-light Image Enhancement. Traditional low-light image enhancement methods, such as gamma correction [RHD*10], histogram equalization [PAA*87], and Retinex-based theory [RJW96], are usually based on the image prior knowledge or algorithms [FLL22]. Reinhard *et al.* [RHD*10] edit the gamma curve of the image, detect the dark and light parts in the image signal, and increase the ratio between them, thereby improving the image contrast effect. Lee *et al.* [LLK13] proposed a contrast enhancement algorithm based on the layered difference representation of 2D histograms, attempting to enhance image contrast by amplifying the grey-level differences between adjacent pixels. Based on the Retinex theory, Jobson *et al.* [JRW97] proposed the best placement of the logarithmic function and Gaussian form to define a specific retinex called SSR to handle grey-world violations. Although these methods can significantly improve image brightness, the enhanced images have poor details and often suffer from severe colour distortion [FLL22]. Since LL-net [LAS17] first utilized deep learning models for low-light enhancement, deep learning methods have become the dominant enhancement method [LGH*21]. Wei *et al.* [WWYL18], applying Retinex theory to deep learning, proposed RetinexNet, which decomposes images into reflectivity and illuminance and enhances illuminance to obtain normal light images. Lv *et al.* [LLWL18] proposed a multi-branch network called MBLEN that was designed to apply enhancement functions through multiple subnets and ultimately generate output images through multi-branch fusion to enhance and denoise low-light images simultaneously. The encoder-decoder structure represented by U-net has also been widely used, and the existing methods have achieved good results through well-designed network structures such as [ZLYF21, FLL22, ZLL22]. Fan *et al.* [FLL22] proposed HWMNet based on an improved hierarchical model M-Net+ and uses a half wavelet attention block on M-Net+ to enrich the features from the wavelet domain. Zhou *et al.* [ZLCL22] proposed LEDNet, consisting of an optical enhancement encoder and a deblurring decoder, and utilizes Filter Adaptive Skip Connections to eliminate ambiguity. Zhang's StableLLVE [ZLYF21] directly uses U-net network validation to increase the effectiveness of optical flow methods. However, the network of this structure often lacks the utilization of global information. Our proposed U-RCSA model emphasizes the guidance of global attention to local features, which are computed to obtain pixel-level attention results through row-column mean and maximum features. We combine deep and shallow information for interaction using subtle residual connection and channel concatenation.

Low-light video enhancement. The biggest challenge facing the transition from low-light image to low-light video enhancement is achieving temporal stability of video frames while ensuring the quality of single-frame enhancement. This means avoiding problems such as flicker and jitter between video frames [XWFJ22a]. Furthermore, obtaining high-quality paired low-light videos is also challenging, as it is necessary to ensure that dynamic scenes under

different lighting conditions are identical. A simple approach is using 3D convolution to train continuous sequences of video frames, extending the image model to a video model. There are also methods to introduce optical flow information into the model's training process, and they use optical flow to describe the motion of objects in dynamic scenes through optical flow. Lai *et al.* [LHW*18] also proposed a deep recurrent network with a ConvLSTM module that uses optical flow as a post-processing method for enforcing temporal consistency in a video. Zhang *et al.* [ZLYF21] use optical flow prior to indicate potential motion from a single image and imitate adjacent frames of images by warping them with corresponding optical flow. They train the model in a twin mode and apply consistency between adjacent frames.

3. Method

As a supervised learning method, given a paired dataset $\langle I_{in}, I_{gt} \rangle$, the U-RCSA model enhances low-light image I_{in} to a normal light image I_{pred} . The process can be formulated as

$$I_{pred} = \mathcal{F}(I_{in}; \theta), \quad (1)$$

where $\mathcal{F}$ represents our model U-RCSA with the trainable parameters θ . We optimize the model by minimizing the error:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \mathcal{L}(I_{pred}, I_{gt}), \quad (2)$$

where $\mathcal{L}$ denotes the loss function. In the model, we propose the RCSA module with a few parameters in Section 3.1 to focus on the global information. Furthermore, we improved the original U-Net, combined it with the RCSA module to form the U-RCSA block, and stacked it three times to form the U-RCSANet in Section 3.2. The network structure is shown in Figure 2. Finally, to ensure the temporal consistency of the method applied to video, we proposed two loss functions in Section 3.3.

3.1. Row-Column Separated Attention module

In the existing research, popular attention modules for image processing, such as ViT [DBK*20] and ViViT [ADH*21], are pixel-to-pixel attention by dividing the image into multiple small patches and then stretching the patches into vectors. ViT is to divide a picture into 16×16 patches of the same size and then stretch each 16×16 patch into a 1×256 vector. However, the number of parameters and the computational effort required for this processing are large. Limited by the image size, this processing is less suitable for image enhancement tasks. Inspired by separable convolution, maxpool and meanpool, we propose a novel attention module based on the row-column mean and maximum. We argue that the mean and maximum values of the rows and columns represent the row-column level information, which is local information relative to the entire feature map. However, the mean and maximum values of all rows and columns can represent the global information of the feature map. The feature map here refers to the input part of the RCSA module, that is, the feature map stitched by the channel from the input and output parts of the previous Improved U-Net module. By self-attention operation, we can fuse all the row-column feature information, thus obtaining the global information of the image.

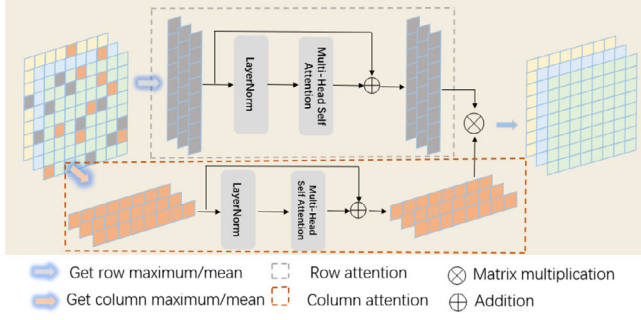


Figure 3: Row-Column Separated Attention module. Average and maximum attention can be obtained through column mean and maximum value.



Figure 4: The results of low light image enhancement for different sizes. From right to left, the image sizes are 1512×1036 , 768×512 , 512×352 and 384×256 .

Figure 3 shows the specific details of the RCSA module. First, we obtain Fh_{avg} and Fw_{avg} for the input feature map $F \in \mathbb{R}^{C \times H \times W}$ by row and column, respectively, and also obtain Fh_{max} and Fw_{max} by row and column, respectively, where the features $Fh_{avg}, Fh_{max} \in \mathbb{R}^{C \times H \times 1}$ and $Fw_{avg}, Fw_{max} \in \mathbb{R}^{C \times 1 \times W}$. We believe that they represent the average features of the feature map rows, the average features of the columns, the significant features of the rows and the significant features of the columns. In this way, when processing the feature maps of the same number of channels, the amount of pixel information processed by our attention module is $(2/H + 2/W)$ of the pixel attention of ViT [DBK*20], which leads to fewer parameters and faster calculation speed. Our attention module applies to images of any size, as shown in Figure 4.

The input of mean and maximum can be calculated as follows

$$Fh_{avg} = \frac{1}{w} \sum_{i=1}^w F(C, H, i), \quad Fh_{max} = \max_i^w F(C, H, i),$$

$$Fw_{avg} = \frac{1}{h} \sum_{i=1}^h F(C, i, W), \quad Fw_{max} = \max_i^h F(C, i, W). \quad (3)$$

Then, the attention's output can be multiplied by the matrix to obtain a dimensional consistent output. This result is obtained by

multiplying the corresponding row and column results, thus ensuring unique differences between pixels.

After completing the matrix multiplication, we obtain two attention results: mean attention and maximum attention. Since they represent the image's average and significant properties, we perform an adaptive weight balancing operation by setting a parameter for each channel separately. The final output can be calculated as

$$F_{out} = a \cdot \Phi(Fh_{avg}) \otimes \Phi(Fw_{avg}) + b \cdot \Phi(Fh_{max}) \otimes \Phi(Fw_{max}), \quad (4)$$

where $a = \text{sigmoid}(\lambda)$, $b = 1 - \text{sigmoid}(\lambda)$, $\otimes$ represents matrix multiplication and $\Phi(x) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$ represents attention operation. Q, K, V are the corresponding query, key and value matrices in the attention mechanism, which are obtained by matrix multiplication of three learnable parameter matrices W_q, W_k and W_v with the input x , λ are adaptive weight parameters.

In order to prove the effectiveness of the RCSA module, we replace our RCSA module with the following modules, which are channel attention module (CAM) [WPLK18], spatial attention module (SAM) [WPLK18], SE attention block from SeNet [HSS18] and ECA attention module from ECANet [WWZ*20]. See Section 4.4 for a comparison of quantitative results.

3.2. U-RCSANet

Improved U-Net. As shown in Figure 2, we perform two down-sampling operations to balance the computational effort while obtaining deep image feature information. After down-sampling, we perform an up-sampling process with a symmetric structure. Unlike U-Net, we do not directly concatenate the channel dimensions of the encoder and decoder [RFB15]. We perform residual and feature re-extraction operations for the encoder output, adding residual structures for each upsampling operation and performing additional feature extraction operations on the encoder results before performing channel cascading operations. After the concatenating operation, we recover the channels by 1×1 convolution layer. That is to say, we fuse shallow and deep features by concatenating the output results of the corresponding layers of the encoder and decoder in the channel dimension and extracting features again from the concatenation results through convolutional layers. Overall, we emphasize the importance of this residual and channel concatenation, which is beneficial in achieving the guidance of shallow information to deep information and can reduce the accumulation of errors caused by model extraction of deep features.

U-RCSA block. The Improved U-Net is connected to the RCSA module as shown in Figure 2. We concatenate the input and output of the Improved U-Net and use the result as the input of the RCSA module. This concatenation is essential and guides shallow information (before re-construction by Improved U-Net) to deep information (after re-construction by Improved U-Net), reducing the accumulation of noise introduced by the model re-construction. The model achieves the interaction of deep and shallow information through subsequent feature extraction operations on the concatenated results, thereby providing guidance for deep information from shallow information. At the end of the attention

module, we use 1×1 convolution to reduce the channel dimension and add a residual connection. Channel concatenation and residual connection between different modules are carefully designed, significantly improving the model's enhancement capabilities. To form U-RCSANet, we stack three U-RCSA blocks whose parameters are the same. The dimensions of channels are elevated and reduced through convolution at both ends of the network.

3.3. Objectives

We have designed two loss functions for different contexts: the image quality loss function and the video loss function.

Image loss function. We adopted the loss function commonly used for low light enhancement and drew on the loss setting scheme in SMNet [LTD*23], combining multiple functions to ensure the model can achieve better results.

- SSIM loss [WBSS04]: The loss is a constraint on structural similarity and helps to produce high-quality results.

$$\mathcal{L}_s = 1 - \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right), \quad (5)$$

where μ_x and μ_y are the mean values of the input and output image, and present variances, σ_{xy} is the covariance and C_1 and C_2 are the constants that keep the denominator from being zero.

- VGG loss [JAFF16]: We use the output features of the VGG [SZ15] network to calculate the perceptual loss. The loss helps the model to fuse the semantic information of the image with the structural information better.

$$\mathcal{L}_p = \|\phi_{ij}(I_{pred}) - \phi_{ij}(I_{gt})\|_1, \quad (6)$$

where ϕ_{ij} describes feature maps computing with j th convolution layer of i th block in VGG-16 network.

- Smooth loss [Gir15]: Compared to traditional L1 or L2 loss, Smooth L1 loss has a smoother curve, which can reduce the training instability caused by noise or outliers and better balance stability and accuracy.

$$\mathcal{L}_{sl1} = \frac{1}{H \cdot W} \sum_{i=1}^H \sum_{j=1}^W \begin{cases} \frac{1}{2}(I_{pred}(i, j) - I_{gt}(i, j))^2, \\ \text{if } |I_{pred}(i, j) - I_{gt}(i, j)| < 1 \\ |I_{pred}(i, j) - I_{gt}(i, j)| - 0.5, \\ \text{if } |I_{pred}(i, j) - I_{gt}(i, j)| \geq 1 \end{cases} \quad (7)$$

- TV loss [SZ15]: Noisy images are often accompanied by high variance [PGC*17], so reducing the total variation between adjacent pixels can help achieve cleaner output.

$$\mathcal{L}_{tv} = \sum_{i=1}^H \sum_{j=1}^W \sqrt{(I(i, j) - I(i+1, j))(I(i, j) - I(i, j+1))}, \quad (8)$$

where $I(i, j)$ represents the pixel at the position (i, j) in the enhanced image.

- MSE loss $\mathcal{L}_{MSE}$: We use MSE loss to regularize the model and improve its robustness.

Video loss function. To ensure the temporal stability of enhanced video, we propose the following video loss functions:

- Neighbouring-frame significant difference loss: Considering that the light and shadow flow in the local area of adjacent frames of dynamic video is not obvious and the object motion amplitude is small, we design the loss function:

$$\mathcal{L}_{dif} = |\text{pool}(\text{bright}_k) - \text{pool}(\text{bright}_{k+1})|, \quad (9)$$

where the $\text{pool}(x)$ represents the maximum pooling operation with a pooling core size of 2 and a step size of 2 and bright_k , bright_{k+1} represent the brightness characteristics of the enhanced neighbouring frames, respectively. Its calculation formula is $\text{bright}_k = I_k^r * 0.299 + I_k^g * 0.587 + I_k^b * 0.114$, where I_k^r , I_k^g , I_k^b represents the R, G and B channels of the k th frame image, respectively. The local flicker phenomenon can be avoided by minimizing the local brightness difference between neighbouring frames.

- Self-similarity loss: Inspired by StableLLVE [ZLYF21], we believe that the neighbouring frame differences of the normal light video should be consistent with the neighbouring frame differences of the enhanced video. This is obvious because it allows the model to learn more about the potential temporal stability information between the two video frames of the input. Therefore, we propose the following loss function.

$$\mathcal{L}_{self} = \|(pred_{k+1} - pred_k) - (gt_{k+1} - gt_k)\|^2, \quad (10)$$

where gt_k and gt_{k+1} are the adjacent frames of normal light video and $pred_k$ and $pred_{k+1}$ are the adjacent frames of the corresponding enhanced video.

Training for image:

$$\begin{cases} \mathcal{L}_{stage1} = \mathcal{L}_p + \mathcal{L}_s + \mathcal{L}_{sl1} + 0.001 * \mathcal{L}_{tv}, \\ \mathcal{L}_{stage2} = \mathcal{L}_p + \mathcal{L}_s + \mathcal{L}_{sl1} + \mathcal{L}_{MSE}, \end{cases} \quad (11)$$

where $\mathcal{L}_{stage1}$ is trained for the first stage to focus on improving the quality of image enhancement, and $\mathcal{L}_{stage2}$ is trained for the second stage to regularize the model and enhance its robustness.

Training for video:

$$\mathcal{L}_{video} = \alpha \cdot \mathcal{L}_{dif} + \beta \cdot \mathcal{L}_{self} + \mathcal{L}_{stage1}, \quad (12)$$

where α and β are the balancing coefficients. We set $\alpha = 2$ and $\beta = 2$.

4. Experiment

To verify the model's validity, we trained and evaluated the model on public image and video datasets, respectively. In Section 4.1, we introduced the datasets and experimental setup. In Sections 4.2 and 4.3, we introduced the quantitative and qualitative experimental results, respectively. Finally, we presented the ablation study of the model in Section 4.4.

4.1. Experimental setup and datasets

Datasets. The LOL dataset consists of 500 pairs of low and normal light images, of which 485 pairs are used for training and 15 pairs for

Table 1: Evaluation results of different methods on the LOL dataset, the bold values indicate the best results. Note that we cannot find an open-source model for experimentation in the blank part of the table.

Methods	Venue	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	NIQE $\downarrow$	RMSE $\downarrow$	DeltaE $\downarrow$
LIME [GLL16]	TIP 2016	16.76	0.56	—	—	—	—
Retinex [WWYL18]	CVPR 2018	16.77	0.56	0.474	8.865	10.09	60.70
MBLLEN [LLWL18]	BMVC 2018	17.56	0.73	0.174	3.585	9.99	56.55
ENlightenGAN [JGL*21]	TIP 2021	17.48	0.65	0.173	4.684	10.07	62.72
MIRNet [ZAK*20]	ECCV 2020	24.14	0.83	0.131	4.201	8.78	28.02
TPET [CLHF22]	EAAI 2022	23.98	0.84	0.122	—	24.46	—
MAXIM [TTZ*22]	CVPR 2022	23.43	0.86	0.099	5.330	9.11	33.06
HWMNet [FLL22]	ICIP 2022	24.24	0.85	0.114	5.141	8.75	29.59
Ours		24.77	0.86	0.097	4.638	8.64	25.49

testing [WWYL18]. MIT Adobe FiveK [BPCD11] is a collection of 5000 RAW format photos taken by a group of photographers using DSLR cameras, covering a wide range of scenes, themes and lighting conditions. SDSD [WXF*21] is the high-quality paired low-light video and normal-light video data collected by professional equipment. It includes SDSD indoor and outdoor data. SDSD indoors has 58 pairs of video data for training and 12 pairs of data for testing, while SDSD outdoors has 67 pairs of video data for training and 13 pairs of video data for testing.

Experimental setup. The source code provided in this paper is implemented using the PyTorch framework and trained using the Adam optimizer, with the parameter betas set as the default value. Betas is the coefficient used to calculate the running average of the gradient and its square, with default values of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. In order to enhance data while reducing training costs, we randomly cropped regions of 128×128 in the LOL image for training. Correspondingly, the regions cropped randomly in the SDSD dataset were 128×240 . All test data obey the dataset corresponding to the paper settings. We set the initial learning rate to $5e-4$, decreasing by 1.2 times every 50 epochs for LOL, every 20 epochs for MIT Adobe FiveK and every 2 epochs for SDSD. LOL trained a total of 600 epochs, and MIT Adobe FiveK trained a total of 100 epochs. Because the SDSD dataset is too large, we only train 10 epochs. We adopted the staged loss function, which improved both metrics, in particular, PSNR by 2%. The metrics of other models in the following experiments are preferentially obtained from published papers. Otherwise, we use the source code provided for experiments.

4.2. Quantitative evaluation

LOL dataset. We compared the U-RCSANet with the latest model on the LOL image dataset. We chose the two most mainstream indicators, PSNR and SSIM. Higher PSNR generally represents less noise, and higher SSIM represents more similar structural properties. Our model exceeds all other models in these two mainstream indicators. In addition, we have also selected LPIPS [ZIE*18], NIQE [SBC12], RMSE and DeltaE [Rob77] to measure the effectiveness of the model comprehensively. LPIPS [ZIE*18] stands for Learned Perceptual Image Patch Similarity, used to measure the similarity between two images based on their perceptual features. A high LPIPS score indicates that the two images being compared

are not similar in their perceptual features. NIQE [SBC12] means Natural Image Quality Evaluator, which measures the quality of a distorted or compressed image. A higher score indicates poorer quality and implies that the distorted or compressed image being evaluated has a lower quality than the original, undistorted image. RMSE represents Root Mean Square Error, and a higher RMSE score means that the two images being compared are more dissimilar. DeltaE [Rob77] is a metric used in colour science to measure the difference between two colours. A lower DeltaE score indicates a smaller colour difference between the two colours. The results are shown in Table 1. It shows that our model is significantly superior to other models in multiple indicators. Our model parameters are fewer than the latest models HWMNet [FLL22] and MAXIM [TTZ*22] in 2022, and only 2% of the HWMNet model, but our quantitative indicators are all ahead of both.

MIT Adobe FiveK dataset. To verify the generality of the method, we trained on the MIT Abode FiveK dataset of 5000 paired datasets. Although the MIT dataset has a wide range of lighting conditions, our model performs well in brightness adaptive enhancement, as shown in Figure 5.

We followed [WZF*19] for our experiments and chose the version of Student C as the ground truth. We followed the specific data partitioning method [WZS*23], with 500 images used for testing and related evaluation metrics results also derived from LLFormer [WZS*23]. We selected classical models as well as the most up-to-date ones. Among them, Uformer [WCB*22a], Restormer [ZAK*22a] and LLformer are all transformer-based models which optimize the transformer mechanism in terms of time complexity. In particular, the LLformer model successively uses row-column attention to reduce the transformer complexity to linear order. The experimental results are shown in Table 2. Compared with the latest model LLformer, although the PSNR of our model does not achieve first place, there is only a 1% difference. However, our SSIM and LPIPS metrics are in the lead. More importantly, the number of parameters in our model is only about 12% of that.

SDSD dataset. We also compared the model with the latest model on the video dataset. For image enhancement quality, we still chose PSNR and SSIM metrics. The results are shown in Table 3. The PSNR of our model is 2% higher than that of DP3DF [XWFJ22a] and 8% higher than that of LSEC [AF22]. We selected AB, MABD, TSSIM and TPSNR common metrics for temporal stability. AB

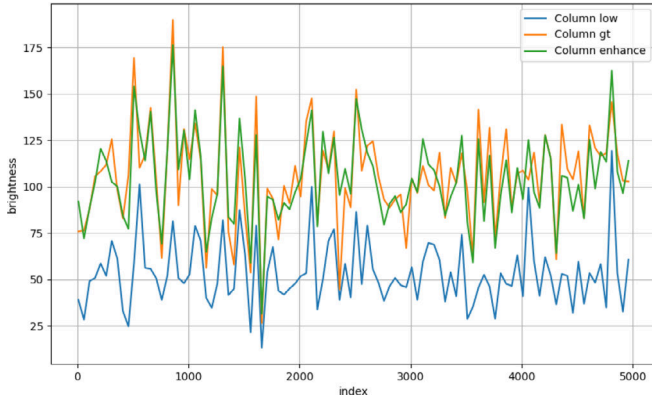


Figure 5: MIT FiveK dataset low light image, normal light image and enhanced image brightness visualization results, it can be found that the brightness of the enhanced image is consistent with the brightness of the normal light image. Through the degree of fluctuation of the three curves can be sent to achieve the model can be adaptive to the low light image brightness enhancement.

Table 2: Evaluation results of different models on MIT Adobe FiveK dataset, the bold values indicate the best results.

Methods	Venue	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
LIME [GLL16]	TIP 2016	13.303	0.750	0.132
Retinex [WWYL18]	CVPR 2018	12.515	0.671	0.254
DSLR [LK21]	TMM 2021	20.244	0.829	0.153
Z_DCE++ [LGL22]	TPAMI 2022	14.611	0.406	0.231
ELGAN [JGL*21]	TIP 2021	17.905	0.836	0.143
Uformer [WCB*22a]	CVPR 2022	21.917	0.871	0.085
Restormer [ZAK*22a]	CVPR 2022	24.923	0.911	0.058
LLFormer [WZS*23]	AAAI 2023	25.753	0.923	0.045
Ours		25.724	0.930	0.041

Table 3: Evaluation results of different models in SDSD dataset, the bold values indicate the best results.

Methods	Venue	PSNR $\uparrow$	SSIM $\uparrow$
DeepUPE [WZF*19]	CVPR 2019	21.82	0.68
MBLLEN [LLWL18]	BMVC 2018	21.79	0.65
SMID [CCDK19]	ICCV 2019	24.09	0.69
SMOID [JZ19b]	ICCV 2019	23.45	0.69
SDSD [WXF*21]	CVPR 2021	24.92	0.73
StableLLVE [ZLYF21]	CVPR 2021	22.85	0.79
LSEC [AF22]	CVPR 2022	25.06	0.84
DP3DF [XWFJ22a]	AAAI 2022	26.46	0.81
Ours		27.01	0.84

metric to measure the average brightness variance between the enhanced video and the ground truth. This metric reflects when the video has unexpected brightness changes or flickers [LLWL18]. Mean absolute brightness differences (MABD) can be viewed as a general level of time derivatives of brightness value on each pixel location [JZ19a]. The mean squared error between the MABD vec-

Table 4: Temporal stability comparison of models on SDSD datasets. The top three results are marked in red, blue and green, respectively.

	AB $\downarrow$	MABD $\downarrow$	TPSNR $\uparrow$	TSSIM $\uparrow$
MBLLEN	0.17	52.36	28.72	0.84
StableLLVE	0.11	11.01	35.02	0.91
w/o $\mathcal{L}_{dif}$ and $\mathcal{L}_{self}$	0.63	16.73	32.47	0.878
w/o $\mathcal{L}_{self}$	0.24	16.83	32.88	0.871
w/o $\mathcal{L}_{dif}$	0.63	17.35	32.00	0.873
Ours	0.26	14.54	33.02	0.886

tor of the enhanced video and the MABD vector of the ground live is the actual MABD index. TPSNR and TSSIM apply the PSNR and SSIM formulas to videos. The calculation process is as follows: first, the PSNR and SSIM indicators between adjacent frames of the video are solved, and then the average of the PSNR indicators of the video is calculated as TPSNR. Therefore, the higher the value, the better. The average of the SSIM indicators of the video is calculated as TSSIM, and the higher the value, the better. The result is shown in Table 4. The first two models are the image-based and optical flow-based video methods, and the last three are the ablation experiments of video loss of our model. It can be found that our temporal stability is greater than that of image-based methods. Although our method is not as good as the method based on optical flow in terms of quantitative indicators comparison, our model has also achieved good results without allowing for calculating optical flow outside the boundary.

4.3. Qualitative evaluation

We show qualitative results on two datasets. Figure 6 shows a qualitative comparison of U-RCSANet with other models on the LOL test set. Generally, with the guidance of global information from the RCSA module, the U-RCSANet can better focus on the global brightness and restore the true colours and details. At the same time, MBLLEN has dark brightness in pictures B and C and distorted colours in A, C and D. The four MAXIM images are all dark, and HWMNet also has a dark problem in pictures A and B, especially the hair colour distortion in the enlarged area in picture C. We composited the test videos after enhancing all the images in the SDSD test dataset. The enhancement results can be seen in the additional materials. Figure 7 shows the enhancement results on the model SDSD dataset. Through comparison, it can be found that there is always a greater difference between the colour of the enhancement results and the ground truth in MBLLEN, and the overall enhancement results of the StableLLVE model are greyer and darker, while our model has better results. According to C in Figure 7, our model can better restore complex details even in very dark areas, while enhanced details by MBLLEN are more blurred and noisy. Figure 8 shows the user preference study. We conducted the user study using the LOL image enhancement results from the test set. Two choices were provided to the users each time. The results are from the 41 user preference choices. All of our models have higher selection rates. We also tested the statistical significance of the users' propensities based on the experimental data. We made the null hypothesis that the probability of our selected model is greater than 50%, 60%,

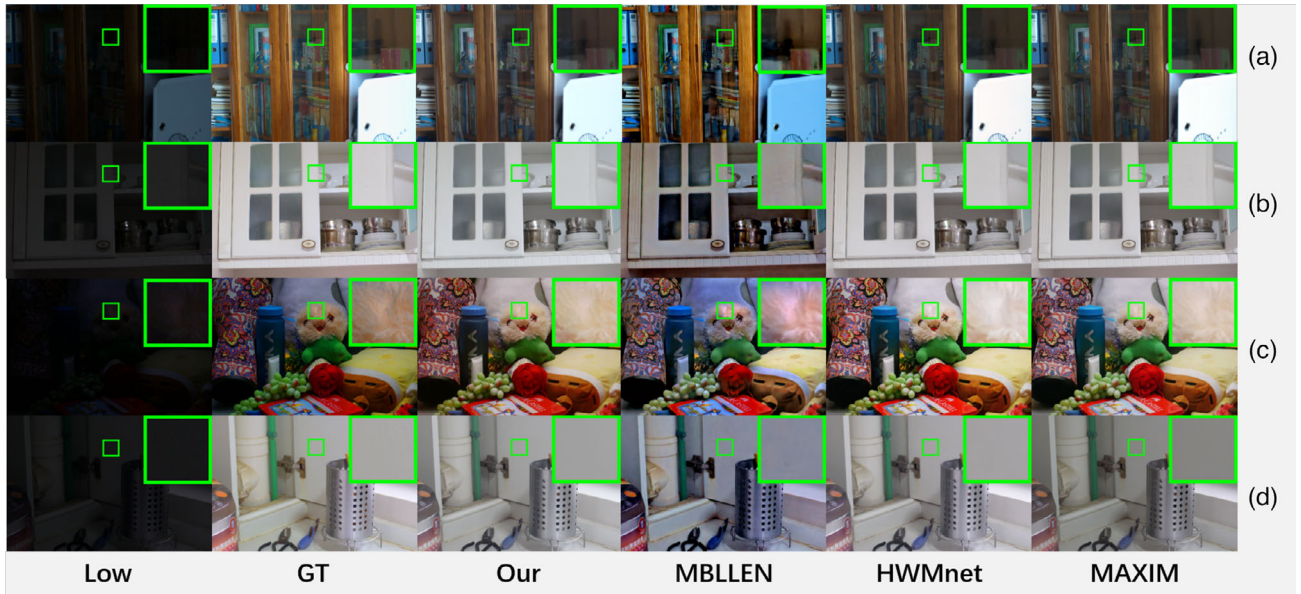


Figure 6: Comparison of enhancement results of different models in LOL dataset. Here, we choose HWMNet, MAXIM and classical methods to compare the overall details.

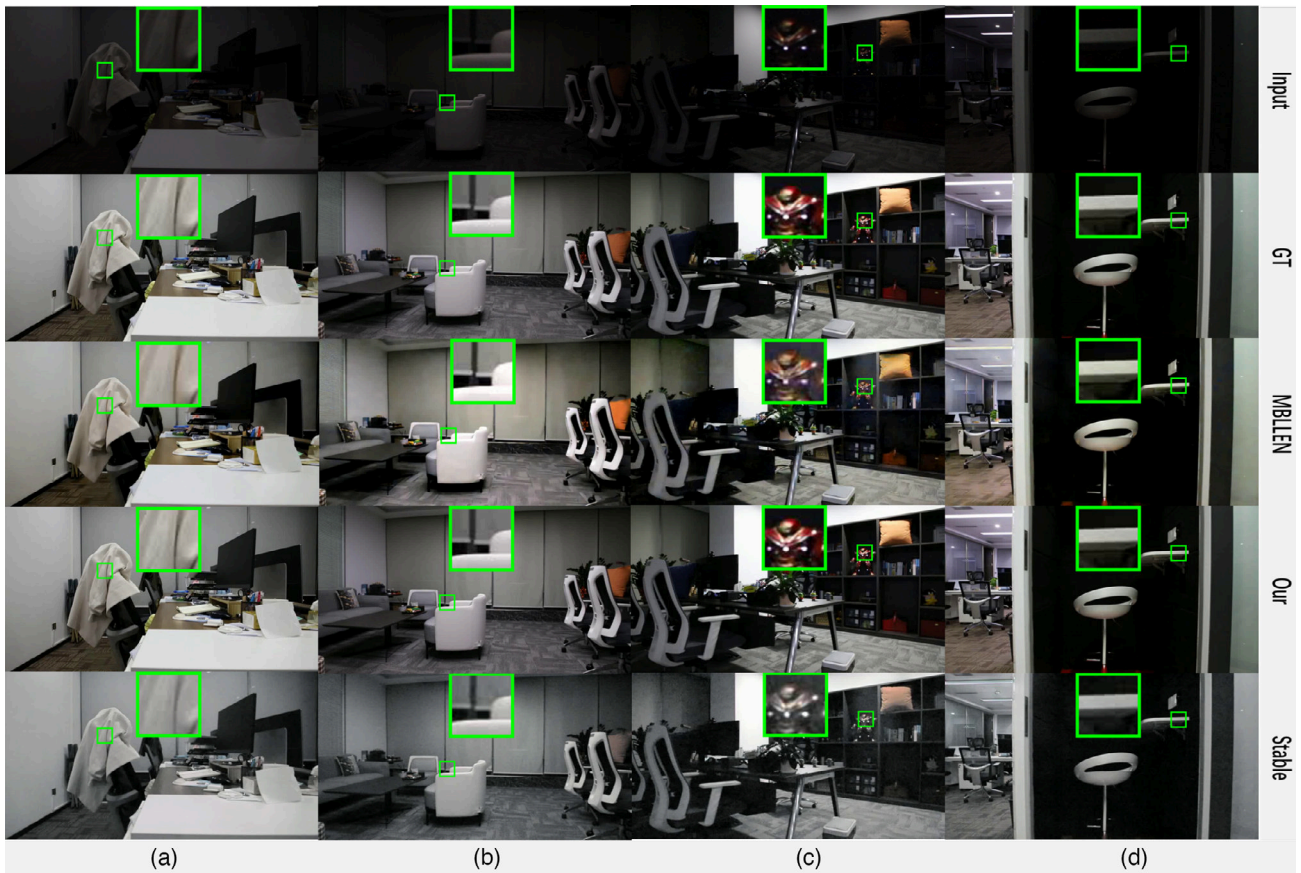


Figure 7: Comparison of enhancement results of different models in the SDSD dataset. Here, we choose MBLLEN and StableLLVE to compare the overall details. Our model can better restore complex details even in very dark areas than other models.

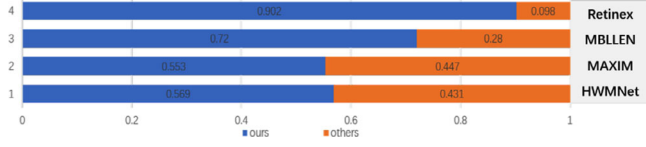


Figure 8: User study. Comparing our model with the two latest models in 2022 and a classical model on LOL dataset, our model has a higher selection rate.

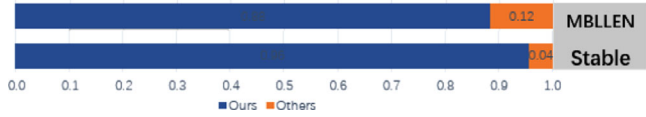


Figure 9: User study. Comparing our model with StableLLVE and MBLLEN on the SDSD video dataset, our model has a higher selection rate.

70% and 90% compared to the other four models, MAXIM, HWMNet, MBLLEN and Retinex, respectively. All the null hypotheses cannot be rejected at a significance level α of 0.05. Figure 9 shows the comparison of video enhancement results of our model with the StableLLVE [ZLYF21] model and the MBLLEN [LLWL18] model. We collected the average selection results of 20 users for 16 videos of SDSD indoor and outdoor data. Although the StableLLVE model outperforms our model in terms of quantitative indicators of temporal consistency, user feedback indicates that neither model has significant flicker issues. Due to the poor image quality enhancement results of the StableLLVE model, our model has a higher selection rate. Although the image quality enhancement results of MBLLEN are better than those of StableLLVE, our model is also more popular due to the serious noise flicker problem of MBLLEN.

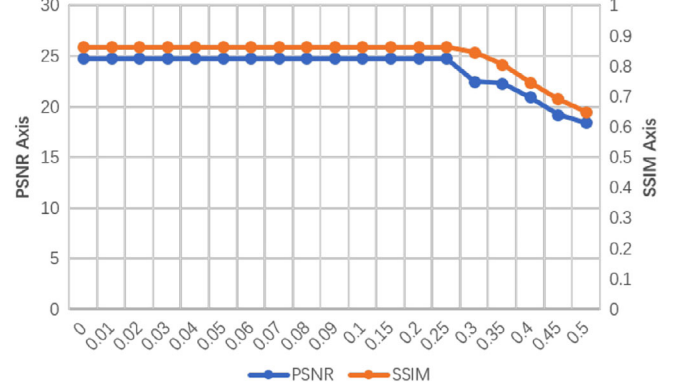


Figure 10: Quantitative noise experiments. We added Gaussian noise after image normalization, and the PSNR and SSIM were evaluated as above. When the standard deviation of the noise exceeds 0.25, the model's performance begins to degrade significantly.

In addition, to verify the robustness of the model to noise, we conducted noise experiments on 15 LOL test set images. First, we normalize the image pixels by dividing them by 255 to normalize them to the $[0,1]$ interval. Then we add Gaussian noise $\mathcal{N}(\mu, \sigma^2)$, $\mu = 0$, $\sigma \in [0, 0.5]$ to the normalized image data, and input the noisy image into the pre-trained model for testing. The experimental results can be seen in Figures 10 and 11, where Figure 10 is a quantitative display and Figure 11 is a qualitative display. For more results, please refer to support files.

4.4. Ablation study

Comparison of the effects of other lightweight attention modules and RCSA attention modules: To verify the effectiveness and computational efficiency of the RCSA module, we replaced

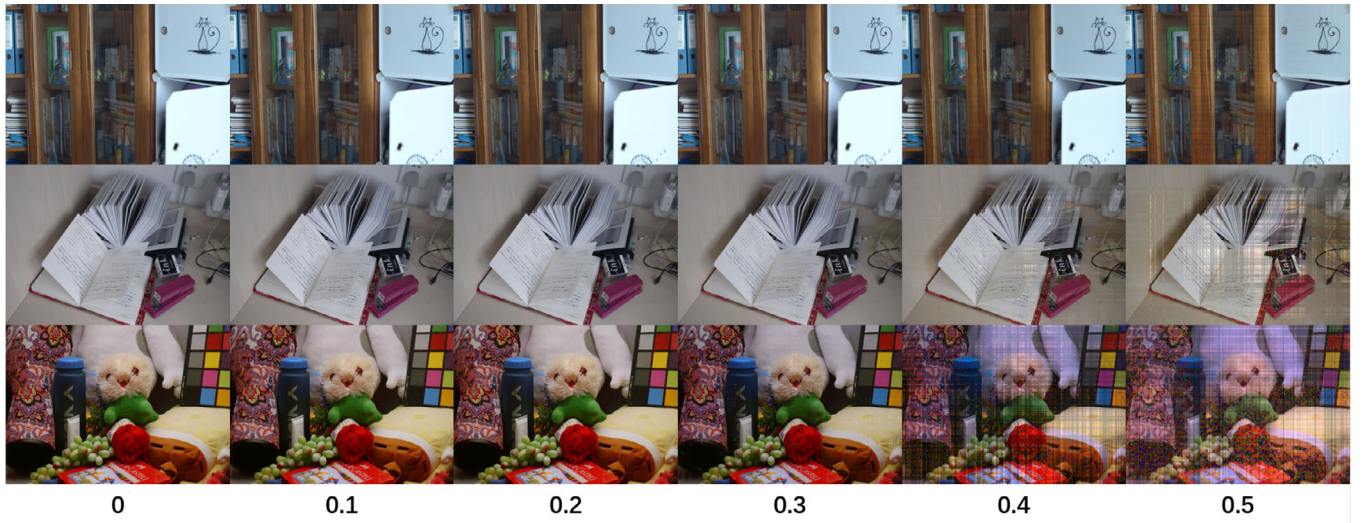


Figure 11: Qualitative noise experiments. We added Gaussian noise after normalizing the image. This figure shows the test results with a standard deviation of 0–0.5. The quality of the results begins to suffer when the standard deviation exceeds 0.3.

Table 5: Quantitative metrics comparison of different attention modules, the bold values indicate the best results. Our RCSA module achieved the best results. Although our model's FPS (frames per second) index is slightly lower than some models, it can be found that the PSNR of other attention module enhancement results is far from the number of our model, and the noise effect has not been minimized as much as possible.

	PSNR ↑	SSIM ↑	LPIPS ↓	FSIM ↑	NIQE ↓	RMSE ↓	DeltaE ↓	FPS ↑
No attention	19.768	0.827	0.129	0.949	4.698	10.02	54.63	13.01
CAM	23.079	0.853	0.107	0.964	4.689	9.25	32.91	9.11
SAM	20.096	0.836	0.128	0.950	4.753	9.73	43.00	12.38
SE block	23.685	0.854	0.115	0.964	4.728	8.86	30.78	12.73
ECA block	23.062	0.850	0.113	0.964	4.657	9.31	42.79	12.75
RCSA(ours)	24.767	0.864	0.097	0.966	4.638	8.64	25.49	11.70

Table 6: Model ablation study on LOL dataset, the bold values indicate the best results. 'Initial U-Net with RCSA' represents the Three U-RCSA block's U-Net module without adding residual connection and convolution reprocessing. 'The three RCSA-U blocks' are the model RCSA-U, which represents swapping the order of the improved Unet and RCSA modules in the original U-RCSA.

Methods	PSNR ↑	SSIM ↑
One U-RCSA block	22.683	0.846
Two U-RCSA blocks	23.468	0.855
Three U-RCSA blocks	24.767	0.864
Three RCSA-U blocks	23.040	0.848
Three U-RCSA_nomax blocks	23.164	0.852
Three U-RCSA_noavg blocks	23.218	0.852
Four U-RCSA blocks	23.865	0.846
Initial U-Net with RCSA	23.160	0.852
Without RCSA module	19.768	0.827

the RCSA module with CAM, SPA, SE block [HSS18] and ECA block [WWZ*20]. These five modules are all lightweight, plug-and-play attention modules. We use the same method to train and test the LOL dataset. The FPS value is the average number of inference frames per second obtained by running 15 LOL test images on an RTX A5000 graphics card 200 times. As shown in Table 5, after adding any attention module, the PSNR and SSIM indicators were greatly improved, and our RCSA module achieved the best results, proving the effectiveness of the RCSA module compared to other lightweight attention modules.

Effectiveness of U-RCSA block: U-RCSA block is the main component of U-RCSANet. We explored the function of this module from the following two aspects. The experimental results were obtained in the LOL dataset and displayed in Table 6. (1) In the aspect of the number of U-RCSA block stacks. Quantitative results are shown in Table 6. The experimental results show that when the number of U-RCSA block stacks gradually increases from one to four, the PSNR and SSIM increase first and then decrease, achieving the best results when the number of stacks is three. Because our modules are not simply stacked, it can be seen from Figure 2 that there is a cross-block residual connection, and the results can be better enhanced through the information interaction of multiple blocks. (2) Whether to improve the U-net. The U-net of the U-RCSA block is our carefully designed and improved structure. Compared with the original U-net structure, the indicators of the improved structure

Table 7: Comparison of parameter quantities and quantitative measures between the model in this paper and the five most advanced models published in 2022 on the LOL dataset, the bold values indicate the best results.

Methods	Venue	Param	PSNR ↑	SSIM ↑
Ours		11.9Mb	24.77	0.864
Ours w/o RCSA		11.5Mb	19.77	0.830
HWMNet [FLL22]	ICIP 2022	762Mb	24.24	0.852
MAXIM [TTZ*22]	CVPR 2022	14.1Mb	23.43	0.863
SNR [XWFJ22b]	CVPR 2022	152.8Mb	24.61	0.842
UNIE [JYT22]	ECCV 2022	512Mb	21.52	0.760
LLformer [WZS*23]	AAAI 2023	282Mb	23.65	0.816

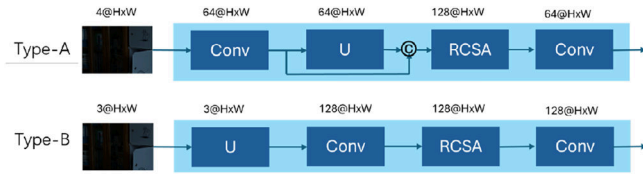
Note: HWMNet, MAXIM, SNR, UNIE and LLFormer all contain their attention modules. In particular, LLformer is a Transformer-based model that reduces the runtime to linear complexity using the row-column attention mechanism. Our model has only one-eighth of its parameters, but the enhancement is superior to it.

significantly rose because we have added shallow information to the deep layer for the original U-net. The residual connection of information avoids the noise amplification caused by over-extracting information. It adds a feature convolution operation for the jump connection between the encoder and decoder to realize the feature conversion between coders better.

Effectiveness of the RCSA module: We conducted the following experiments using the mean-attention and maximum-attention modules of the RCSA structure. When the RCSA module is not used, the two indicators drop significantly because the network lacks the extraction and utilization of global information after removing RCSA and cannot guide local areas due to large local noise. When we only use the mean or maximum value, the effect of the model also decreases because the mean and maximum values of the row and column represent different feature information of the row and column. Therefore, the adaptive accumulation of the mean and maximum matrices can avoid the excessive loss of information down-sampling while utilizing different features as much as possible. Table 7 compares our model with the five state-of-the-art models with attention operations in 2022 regarding parameter quantities and quantitative metrics on the LOL dataset. Compared with other attention-related models in 2022, our model is smaller and has the best effect. After adding the RCSA module, the number of our parameters only increased by 3%. The model's size increased by 0.4Mb, but the PSNR and SSIM improved by 25% and 4%, respectively.

Table 8: Comparison of various transformer-based structures, the bold values indicate the best results.

Methods	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Uformer [WCB*22b]	18.547	0.721	0.321
Restormer [ZAK*22b]	22.365	0.816	0.141
LLformer [WZS*23]	23.649	0.816	0.169
Retinexformer [CBL*23]	25.160	0.845	0.131
U-RWKV	23.051	0.855	0.095
U-Restormer	23.900	0.856	0.101
RCSA-U	23.040	0.848	0.109
U-RCSA(Ours)	24.767	0.864	0.097

**Figure 12:** Two ways of fusion RCSA with Uformer. The blue area is a basic block. If it is a cascade model, there will be three blocks with shared parameters. The $\oplus$ represents channel concatenation, the $x@H \times W$ above each component represents the output dimension of the component, and U represents the Unet or Uformer models. This combination is closest to our model. Type-A's input image has four channels, which means a brightness channel is added to the image.

Comparison of the network structures: In order to verify the effectiveness of U-RCSA, a combination of UNet local convolution and RCSA global attention, we try to use the transformer structure and adjust the order of RCSA modules for comparison experiments. The results are shown in Table 8. Uformer, Restormer and LLformer are the UNet-based network models that replace the local convolution module with the transformer module. Restormer achieves linear complexity by computing self-attention in the channel dimension through the multi-Dconv head ‘transposed’ attention (MDTA) module. LLFormer significantly reduces linear complexity by the axis-based multi-head self-attention and cross-layer attention fusion block. In Table 8, U-Restormer is to substitute the RCSA module with Restormer’s MDTA with linear complexity. U-RWKV is to replace the RCSA module with the RWKV [PAA*23] model, where the RWKV model is a linear attention model that combines RNN and Transformer. The modified structure model follows the same training method on the LOL dataset.

In addition, we tested the ability of RCSA on Unet and Uformer and attempted two fusion structures, as shown in Figure 12. Type-A is closer to our model structure, and Type-B preserves the original structure of Uformer and Unet. After combining the RCSA module with Uformer and Unet, the results of both methods were improved, as shown in Table 9, and the fusion method closer to the model in this paper had better results, proving the RCSA module’s role. However, we found that Uformer and Unet may not be suitable for the cascade structure of this paper, and the effect of the Type-A fusion method was reduced after cascading.

Table 9: Uformer/Unet combined with RCSA and cascade experiments. Cascade means the same three-time parameter sharing as in the article.

Methods	PSNR $\uparrow$	SSIM $\uparrow$
Type-A Unet with RCSA (cascade)	23.160	0.852
Type-A Unet with RCSA	23.483	0.855
Type-B Unet with RCSA (cascade)	21.873	0.829
Type-B Unet with RCSA	22.370	0.845
Type-A Uformer with RCSA (cascade)	24.254	0.859
Type-A Uformer with RCSA	24.470	0.859
Type-B Uformer with RCSA (cascade)	23.621	0.857
Type-B Uformer with RCSA	22.668	0.860

**Figure 13:** Failure cases in extreme environments.

5. Limitation

Although numerous experiments have shown that our method can achieve good results in most cases, when the input image is too dark or noisy, the enhancement effect of our model will deteriorate. Figure 13 shows a city night scene image we collected. It can be observed that the light source in the image is mixed, and some areas are too bright and too dark, as well as a large amount of noise. Moreover, due to the long shooting distance, the image itself is also relatively blurry. The green box in the figure has a moderate enhancement effect due to the influence of scattered light sources, while the blue box in the figure has a poor enhancement effect due to the area being too dark and the overall noise being high. In addition, although our model does not perform simple brightness enhancement on low-light images, it learns how to enhance low-light images into normal-light images that conform to human vision and achieve success in low-light image enhancement under different lighting conditions and scenes. Achieved good results, as shown in Figure 5. However, we still found that when the input is a normal light image, our model still has a certain degree of enhancement, reducing human vision’s aesthetics. This will also be an improvement direction in the future, such as by adding similar training data or relevant constraints.

6. Discussion

The results in Table 4 indicate that our temporal stability metrics exhibit a performance that lags behind those derived from optical flow techniques. At the same time, our model only relies on two adjacent frames of images. However, our proposed loss function still facilitates the model in learning the time stability information between video frames. Therefore, the $\mathcal{L}_{dif}$ and $\mathcal{L}_{self}$ loss functions are an

Table 10: Test results of different models on the LSRW dataset, the bold values indicate the best results. The first three models were tested after they had been trained on LSRW. The following three models were directly tested on the LSRW test after they had already been trained on the LOL dataset.

Methods	Venue	PSNR $\uparrow$	SSIM $\uparrow$	NIQE $\downarrow$
Retinex [WWYL18]	CVPR 2018	15.48	0.347	10.31
KinD [ZGM*21]	IJCV 2021	16.41	0.476	11.13
URetinex [WWZ*22]	CVPR 2022	18.10	0.514	10.76
SMNet [LTD*23]	TMM 2023	17.59	0.521	3.97
LLFormer [WZS*23]	AAAI 2023	18.13	0.519	3.73
Ours		17.97	0.513	3.86

idea and attempt to constrain the consistency between video frames. They can also be used in optical flow-based methods to help the model better constrain the inter-frame loss. Among them, the loss $\mathcal{L}_{dif}$ focuses on the constraints on the temporal consistency of the video itself, which is a universal temporal constraint. Imagine that if the input video frame has temporal inconsistency, there could be significant local brightness differences leading to a significant loss of $\mathcal{L}_{dif}$, which will force the model to minimize the local brightness differences of the enhancement results as much as possible to achieve better temporal consistency and make the model more robust to the data. The loss $\mathcal{L}_{self}$ focuses on ensuring that the enhanced video is as close to the real video as possible, allowing the model to learn the temporal consistency of the training video's characteristics. In addition, our model only constrains the temporal consistency of video frames from the perspective of loss. A better approach would be to fuse attention and feature information between different frames, which will also be our future direction of effort.

To verify the model's generalization ability, we conducted the following generalization test experiments on LSRW dataset [HXY*23], and the results are shown in Table 10. SMNet, LLFormer and ours obtained the results of pre-trained models from the LOL dataset tested directly on LSRW without fine-tuning. Retinex [WWYL18], KinD [ZGM*21] and URetinex [WWZ*22] got the test results after training on the LSRW dataset. Although our model did not take the lead in the generalization experiments, the gap with the highest score was very small and similar to the test results of the pre-trained model. This proved to be a good generalization effect of our model.

7. Conclusion

We have modified the U-Net architecture to achieve a better fusion of deep and shallow information and proposed an RCSA module with a small number of parameters. Our method of taking row–column information as the basic unit of attention input can make good use of global information, and pixel-level attention can be obtained through the fusion of row–column attention results. The RCSA module can be easily migrated to other models as a lightweight module. Meanwhile, we set the loss functions for the features of adjacent frames, which guarantee the temporal stability of video frames. The results of a large number of comparative experiments on image and video datasets demonstrate the effectiveness of our model.

Acknowledgements

This work is supported in part by the Beijing Natural Science Foundation under No. L221013 and the National Natural Science Foundation of China under Grant Nos. 62102162 and 62203184.

Conflicts of Interest

We declare that they have no conflicts of interest.

References

- [ADH*21] ARNAB A., DEGHANI M., HEIGOLD G., SUN C., LUČIĆ M., SCHMID C.: ViViT: A video vision transformer. In *ICCV* (Montreal, BC, Canada, 2021), pp. 6836–6846. <https://doi.org/10.1109/ICCV48922.2021.00676>.
- [AF22] AGHAJANZADEH S., FORSYTH D.: Long scale error control in low light image and video enhancement using equivariance. *arXiv preprint arXiv:2206.01334* (2022). <https://doi.org/10.48550/arXiv.2206.01334>.
- [BPCD11] BYCHKOVSKY V., PARIS S., CHAN E., DURAND F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011* (Colorado Springs, 2011), pp. 97–104. <https://doi.org/10.1109/CVPR.2011.5995413>.
- [CBL*23] CAI Y., BIAN H., LIN J., WANG H., TIMOFTE R., ZHANG Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Paris, France, 2023), pp. 12504–12513. <https://arxiv.org/abs/2303.06705>.
- [CCDK19] CHEN C., CHEN Q., DO M. N., KOLTUN V.: Seeing motion in the dark. In *ICCV* (2019), COEX Convention Center, pp. 3185–3194. <https://doi.org/10.1109/ICCV.2019.00328>.
- [CLHF22] CUI H., LI J., HUA Z., FAN L.: TPET: Two-stage perceptual enhancement transformer network for low-light image enhancement. *Engineering Applications of Artificial Intelligence* 116 (2022), 105411. <https://doi.org/10.1016/j.engappai.2022.105411>.
- [DBK*20] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISENBORN D., ZHAI X., UNTERTHINER T., DEGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J., HOULSBY N.: An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (Vienna, Austria, 2021). <https://arxiv.org/abs/2010.11929>.
- [FLL22] FAN C.-M., LIU T.-J., LIU K.-H.: Half wavelet attention on M-Net+ for low-light image enhancement. In *2022 IEEE International Conference on Image Processing (ICIP)* (Bordeaux, France, 2022), pp. 3878–3882. <https://doi.org/10.1109/ICIP46576.2022.9897503>.
- [Gir15] GIRSHICK R.: Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 1440–1448. <https://arxiv.org/abs/1504.08083>.

- [GLL16] GUO X., LI Y., LING H.: LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing* 26, 2 (2016), 982–993. <https://doi.org/10.1109/TIP.2016.2639450>.
- [HDC*22] HAN J., DENG X., CAI X., YANG Z., XU H., XU C., LIANG X.: Laneformer: Object-aware row-column transformers for lane detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Virginia, 2022), pp. 799–807. <https://doi.org/10.48550/arXiv.2203.09830>.
- [HSS18] HU J., SHEN L., SUN G.: Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, Utah, 2018), pp. 7132–7141. <https://doi.org/10.1109/tpami.2019.2913372>.
- [HXY*23] HAI J., XUAN Z., YANG R., HAO Y., ZOU F., LIN F., HAN S.: R2RNet: Low-light image enhancement via real-low to real-normal network. *Journal of Visual Communication and Image Representation* 90 (2023), 103712. <https://doi.org/10.1016/j.jvcir.2022.103712>.
- [JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *ECCV* (Amsterdam, the Netherlands, 2016), Springer, pp. 694–711. <https://arxiv.org/abs/1603.08155>.
- [JGL*21] 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 Transactions on Image Processing* 30 (2021), 2340–2349. <https://doi.org/10.1109/TIP.2021.3051462>.
- [JRW97] JOHNSON D. J., RAHMAN Z.-U., WOODDELL G. A.: Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing* 6, 3 (1997), 451–462. <https://doi.org/10.1109/83.557356>.
- [JYT22] JIN Y., YANG W., TAN R. T.: Unsupervised night image enhancement: When layer decomposition meets light-effects suppression. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII* (Tel-Aviv, 2022), Springer, pp. 404–421. <https://doi.org/10.48550/arXiv.2207.10564>.
- [JZ19a] JIANG H., ZHENG Y.: Learning to see moving objects in the dark. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul, Korea, 2019), pp. 7323–7332. <https://doi.org/10.1109/ICCV.2019.00742>.
- [JZ19b] JIANG H., ZHENG Y.: Learning to see moving objects in the dark. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), COEX Convention Center, pp. 7323–7332. <https://doi.org/10.1109/ICCV.2019.00742>.
- [KYK*17] KO S., YU S., KANG W., PARK C., LEE S., PAIK J.: Artifact-free low-light video enhancement using temporal similarity and guide map. *IEEE Transactions on Industrial Electronics* 64, 8 (2017), 6392–6401.
- [LAS17] LORE K. G., AKINTAYO A., SARKAR S.: LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition* 61 (2017), 650–662. <https://doi.org/10.1016/j.patcog.2016.06.008>.
- [LGH*21] LI C., GUO C., HAN L.-H., JIANG J., CHENG M.-M., GU J., LOY C. C.: Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2021), 9396–9412. <https://arxiv.org/abs/2104.10729>.
- [LGL22] LI C., GUO C., LOY C. C.: Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 8 (2022), 4225–4238. <https://doi.org/10.1109/TPAMI.2021.3063604>.
- [LHW*18] LAI W.-S., HUANG J.-B., WANG O., SHECHTMAN E., YUMER E., YANG M.-H.: Learning blind video temporal consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany, 2018), pp. 170–185. https://doi.org/10.1007/978-3-030-01267-0_11.
- [LHWH22] LU B., HU Q., WANG Y., HU G.: RCANet: Row-column attention network for semantic segmentation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (Singapore, 2022), IEEE, pp. 2604–2608. <https://doi.org/10.1109/ICASSP43922.2022.9746869>.
- [LK21] LIM S., KIM W.: DSLR: Deep stacked Laplacian restorer for low-light image enhancement. *IEEE Transactions on Multimedia* 23 (2021), 4272–4284. <https://doi.org/10.1109/TMM.2020.3039361>.
- [LLK13] LEE C., LEE C., KIM C.-S.: Contrast enhancement based on layered difference representation of 2D histograms. *IEEE Transactions on Image Processing* 22, 12 (2013), 5372–5384. <https://doi.org/10.1109/TIP.2013.2284059>.
- [LLWL18] LV F., LU F., WU J., LIM C.: MBLLEN: Low-light image/video enhancement using CNNs. In *BMVC* (Newcastle, UK, 2018), pp. 4. <https://phi-ai.buaa.edu.cn/project/MBLLEN/>.
- [LTD*23] LIN S., TANG F., DONG W., PAN X., XU C.: SMNet: Synchronous multi-scale low light enhancement network with local and global concern. *IEEE Transactions on Multimedia* (2023). <https://doi.org/10.1109/83.557356>.
- [PAA*87] PIZER S. M., AMBURN E. P., AUSTIN J. D., CROMARTIE R., GESELOWITZ A., GREER T., TER HAAR ROMENY B., ZIMMERMAN J. B., ZUIDERVELD K.: Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing* 39, 3 (1987), 355–368. [https://doi.org/10.1016/S0734-189X\(87\)80186-X](https://doi.org/10.1016/S0734-189X(87)80186-X).
- [PAA*23] PENG B., ALCAIDE E., ANTHONY Q., ALBALAK A., ARCADINHO S., BIDERMAN S., CAO H., CHENG X., CHUNG M., GRELLA M., GV K. K., HE X., HOU H., LIN J., KAZIENKO P., KOCAN J., KONG J., KOPTYRA B., LAU H., MANTRI K. S. I., MOM F., SAITO A., SONG G., TANG X., WANG B., WIND J. S., WOZNIAC S.,

- ZHANG R., ZHANG Z., ZHAO Q., ZHOU P., ZHOU Q., ZHU J., ZHU R.-J.: RWKV: Reinventing RNNs for the transformer era. In *The 2023 Conference on Empirical Methods in Natural Language Processing* (Singapore, 2023). <https://arxiv.org/abs/2305.13048> (2023), arXiv:2305.13048.
- [PGC*17] PASZKE A., GROSS S., CHINTALA S., CHANAN G., YANG E., DEVITO Z., LIN Z., DESMAISON A., ANTIGA L., LERER A.: Automatic differentiation in PyTorch (2017). <https://openreview.net/forum?id=BJJsrnfCZ>.
- [RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (2015), Springer, pp. 234–241. <https://arxiv.org/abs/1505.04597>.
- [RHD*10] REINHARD E., HEIDRICH W., DEBEVEC P., PATTANAİK S., WARD G., MYSKOWSKI K.: *High Dynamic Range Imaging: Acquisition, Display, and Image-based Lighting*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2010. <https://dl.acm.org/doi/10.5555/1208706>.
- [RJW96] RAHMAN Z.-U., JOBSON D. J., WOODDELL G. A.: Multi-scale retinex for color image enhancement. In *Proceedings of 3rd IEEE International Conference on Image Processing* (1996), vol. 3, IEEE, pp. 1003–1006. <https://doi.org/10.1109/ICIP.1996.560995>.
- [Rob77] ROBERTSON A. R.: The CIE 1976 color-difference formulae. *Color Research & Application* 2, 1 (1977), 7–11. <https://doi.org/10.1002/j.1520-6378.1977.tb00104.x>.
- [SBC12] SAAD M. A., BOVIK A. C., CHARRIER C.: Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing* 21, 8 (2012), 3339–3352. <https://doi.org/10.1109/TIP.2012.2191563>.
- [SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations (ICLR 2015)* (San Diego, CA, USA, 2015), pp. 1–14. <https://arxiv.org/abs/1409.1556>.
- [TTZ*22] TU Z., TALEBI H., ZHANG H., YANG F., MILANFAR P., BOVIK A., LI Y.: MAXIM: Multi-axis MLP for image processing. In *CVPR* (New Orleans, 2022), pp. 5769–5780. <https://doi.org/10.48550/arXiv.2201.02973>.
- [TXB19] TENG L., XUE F., BAI Q.: Remote sensing image enhancement via edge-preserving multiscale retinex. *IEEE Photonics Journal* 11, 2 (2019), 1–10. <https://doi.org/10.1109/JPHOT.2019.2902959>.
- [WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. <https://doi.org/10.1109/TIP.2003.819861>.
- [WCB*22a] WANG Z., CUN X., BAO J., ZHOU W., LIU J., LI H.: Uformer: A general u-shaped transformer for image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, 2022), pp. 17662–17672. <https://doi.org/10.1109/CVPR52688.2022.01716>.
- [WCB*22b] WANG Z., CUN X., BAO J., ZHOU W., LIU J., LI H.: Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, June 2022), pp. 17683–17693. <https://doi.org/10.48550/arXiv.2106.03106>.
- [WPLK18] WOO S., PARK J., LEE J.-Y., KWEON I. S.: CBAM: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich, Germany, September 2018). <https://arxiv.org/abs/1807.06521>.
- [WWYL18] WEI C., WANG W., YANG W., LIU J.: Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference* (Northumbria University, English 2018). <https://doi.org/10.48550/arXiv.1808.04560>.
- [WWZ*20] WANG Q., WU B., ZHU P., LI P., ZUO W., HU Q.: ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, Washington, 2020), pp. 11534–11542. <https://doi.org/10.1109/CVPR42600.2020.01155>.
- [WWZ*22] WU W., WENG J., ZHANG P., WANG X., YANG W., JIANG J.: 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), pp. 5901–5910. <https://doi.org/10.1109/CVPR52688.2022.00581>.
- [WXF*21] WANG R., XU X., FU C.-W., LU J., YU B., JIA J.: Seeing dynamic scene in the dark: A high-quality video dataset with mechatronic alignment. In *ICCV* (Montreal, BC, Canada, 2021), pp. 9700–9709. <https://doi.org/10.1109/ICCV48922.2021.00956>.
- [WZF*19] WANG R., ZHANG Q., FU C.-W., SHEN X., ZHENG W.-S., JIA J.: Underexposed photo enhancement using deep illumination estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (California, 2019), pp. 6842–6850. <https://doi.org/10.1109/CVPR.2019.00701>.
- [WZHL13] WANG S., ZHENG J., HU H.-M., LI B.: Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing* 22, 9 (2013), 3538–3548. <https://doi.org/10.1109/TIP.2013.2261309>.
- [WZS*23] WANG T., ZHANG K., SHEN T., LUO W., STENGER B., LU T.: Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington DC, 2023), pp. 2654–2662. <https://doi.org/10.48550/arXiv.2212.11548>.

- [XWFJ22a] XU X., WANG R., FU C.-W., JIA J.: Deep parametric 3D filters for joint video denoising and illumination enhancement in video super resolution. *Proceedings of the AAAI Conference on Artificial Intelligence* (Washington, 2023), 37, 3, 3054–3062. <https://doi.org/10.1609/aaai.v37i3.25409>.
- [XWFJ22b] XU X., WANG R., FU C.-W., JIA J.: SNR-aware low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, 2022), pp. 17714–17724. <https://doi.org/10.1109/CVPR52688.2022.01719>.
- [XZC*22] XU H., ZHANG J., CAI J., REZATOFIGHI H., TAO D.: GM-Flow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, Louisiana, 2022), pp. 8121–8130. <https://doi.org/10.1109/CVPR52688.2022.00795>.
- [ZAK*20] ZAMIR S. W., ARORA A., KHAN S., HAYAT M., KHAN F. S., YANG M.-H., SHAO L.: Learning enriched features for real image restoration and enhancement. In *ECCV* (Glasgow, UK, 2020), Springer, pp. 492–511. <https://arxiv.org/abs/2003.06792>.
- [ZAK*22a] ZAMIR S. W., ARORA A., KHAN S., HAYAT M., KHAN F. S., YANG M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, June 2022), pp. 5728–5739. <https://doi.org/10.1109/CVPR52688.2022.00564>.
- [ZAK*22b] ZAMIR S. W., ARORA A., KHAN S., HAYAT M., KHAN F. S., YANG M.-H.: Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (New Orleans, 2022), pp. 5728–5739. <https://doi.org/10.48550/arXiv.2111.09881>.
- [ZGM*21] ZHANG Y., GUO X., MA J., LIU W., ZHANG J.: Beyond brightening low-light images. *International Journal of Computer Vision* 129 (2021), 1013–1037. <https://doi.org/10.1007/s11263-020-01407-x>.
- [ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, Utah, 2018), pp. 586–595. <https://doi.org/10.1109/CVPR.2018.00068>.
- [ZJJ*21] ZHANG Z., JIANG Y., JIANG J., WANG X., LUO P., GU J.: STAR: A structure-aware lightweight transformer for real-time image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 4106–4115. <https://doi.org/10.1109/ICCV48922.2021.00407>.
- [ZLCL22] ZHOU S., LI C., CHANGE LOY C.: LEDNet: Joint low-light enhancement and deblurring in the dark. In *ECCV* (Tel Aviv, 2022), Springer, pp. 573–589.
- [ZLL22] ZHOU S., LI C., LOY C. C.: LEDNet: Joint low-light enhancement and deblurring in the dark. In *ECCV* (Tel Aviv, Israel, 2022). <https://doi.org/10.48550/arXiv.2202.03373>.
- [ZLYF21] ZHANG F., LI Y., YOU S., FU Y.: Learning temporal consistency for low light video enhancement from single images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, 2021), pp. 4965–4974. <https://doi.org/10.1109/CVPR46437.2021.00493>.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information