Multimedia Tools and Applications (2023) 82:15271–15286
https://doi.org/10.1007/s11042-022-13830-y

Effective low-light image enhancement with multiscale and context learning network

Qiao Li[1] · Bin Jiang[1] · Xiaochen Bo[2] · Chao Yang[1] · Xu Wu[3]

Received: 9 May 2022 / Revised: 18 August 2022 / Accepted: 6 September 2022 /
Published online: 3 October 2022

Abstract

Convolutional Neural Network (CNN) has been widely used in low-light image enhancement task, and has achieved good enhancement results. However, the enhancement results not only are limited by the convolution kernel, but also are affected by the different shapes and sizes of low-light regions. CNN can only capture local dependencies. It is difficult to obtain long-distance dependencies and multiscale features from images, resulting in over/under enhancement. To alleviate these problems, we propose a Multiscale and Context Learning Network (MCLNet) for adaptive low-light enhancement by multiscale feature extraction and global relationships learning. Concretely, in order to obtain discriminative representation in diverse low-light regions, we design an Attentive Residual Multiscale Block (ARMB) to captures valuable multiscale features through spatial attention mechanisms at different scales. Further, we propose a Bottleblock of Scale Aggregation Module (BSAM) to learn hierarchical discriminative features based on th ARBM. Finally, to further adaptive enhancement from globel view, we present a Context Encoding Module (CEM) to model long-distance dependencies by Transformer. Experimental results show that our proposed MCLNet achieves superior performance of low-light images enhancement than some state-of-the-art methods.

✉ Bin Jiang
jiangbin@hnu.edu.cn

Qiao Li
hliqiao@hnu.edu.cn

Xiaochen Bo
boxiaoc@163.com

Chao Yang
yangchaoedu@hnu.edu.cn

Xu Wu
csxuwu@163.com

[1] Hunan University, ChangSha, China

[2] Beijing Institute of Radiation Medicine, Beijing, China

[3] Shenzhen University, Shenzhen, China

Keywords Low-light image enhancement · Transformer · CNN

1 Introduction

In recent years, people's requirements for the quality of image enhancement have become higher and higher, and more attention has been paid to the problem of over/under image enhancement. At the same time, they also pay attention to the detailed information on image enhancement. However, images in low-light conditions often have degradation phenomena, such as low contrast, low brightness, and severe noise. These degraded images pose challenges to various fundamental tasks in computer vision, such as semantic segmentation, object detection, and tracking. Therefore, it is significant to develop effective image enhancement methods for obtaining high-quality images from degraded inputs.

Current methods for solving low-light image enhancement task can be divided into two categories: traditional methods and learning-based methods. Traditional methods are mostly based on Histogram Equalization (HE) [11, 35] and Retinex [15, 17]. Concretely, the HE-based methods can effectively improve image brightness and gain enhancement effects for images with low contrast, but they easily amplify noise, leading to artifacts [21, 32]. The Retinex-based methods [6, 17] can not only restore the illumination but also improve brightness and contrast of low-light image. However, it is difficult to accurately estimate and process the illumination, resulting in over enhancement in brighter regions and color distortion [33, 38]. Fortunately, the CNN has shown impressive results in many computer vision tasks. Using the attention mechanism [44, 50] and contextual information, CNN can generate attention-perception from the original image and extract multiscale features [18, 46]. Driven by these results, CNN-based low-light image enhancement methods have been continuously developed. For example, the adaptive low-light image enhancement framework [20] based CNN greatly enhances image contrast, color, and detail information. However, most of the existing CNN-based methods mainly focus on the restoration of image brightness, texture, and color [43]. Because of uneven local illumination, serious loss of color information and detail information, it is prone to over enhancement or under enhancement problems.

To this end, we propose a novel Multiscale and Context Learning Network (MCLNet) for adaptive low-light enhancement. MCLNet consists of Efficient Feature Extraction Subnetwork (EFES) and Upsampling Subnetwork (US). Concretely, EFES consists of two parts: a Bottleblock of Scale Aggregation Module (BSAM) and a Context Encoding Module (CEM). They respectively learn multi-scale representation and model global dependencies. By stacking the BSAM, the proposed model can obtain the hierarchical features from low-light images. The core part of BSAM is an Attentive Residual Multiscale Block (ARMB). The ARMB fuses the different scale features and weights these features through the channel attention mechanism. It allows our model to enhance multiscale learning ability and highlight valuable information. Then, we apply CEM to capture long-range dependencies, which is performed by Transformer. Furthermore, to fully extract the texture and details information of the network, we introduce a perceptual loss function to guide the network to generate high-quality images.

In summary, the main contributions of MCLNet are as follows:

1. From both local and global perspective, we propose a Multiscale and Context Learning Network (MCLNet) for adaptive low-light enhancement by multiscale representation learning and long-range dependencies modeling.

2. We introduce a multiscale learning block with residual learning and attention mechanisms, called ARMB, to learn multiscale representations. Based on the ARMB, the introduced BSAM captures information about the global and local contexts, enabling the network to enhance local regions based on global information.
3. Our proposed CEM can model global relationships from multiscale information in diverse low-light regions, thereby enhancing the adaptive low-light enhancement ability.
4. Extensive experiments show that our proposed MCLNet achieves state-of-the-art results, demonstrating its effectiveness in low-light image enhancement. As a side contribution, we have released the models and codes.[1]

The remainder of this article is arranged as follows: Section 2 introduces related works. Section 3 presents the proposed MCLNet in detail. Section 4 describes experiments and discusses the results of different methods.

2 Related works

2.1 Traditional low-light enhancement approaches

Traditional low-light enhancement methods include HE-based methods [11, 35] and Retinex-based methods [15, 17]. The HE-based methods enhance images by stretching the dynamic range of the image. Zhuang et al. [53] propose an image enhancement method based on entropy-based adaptive sub-histogram equalization. Mun et al. [26] design a new edge-enhanced double histogram equalization method using guided image filters. These methods are simple and efficient, but their enhanced results are easily impacted by artifacts leading to low visual pleasing.

The Retinex-based methods simulate the human visual system, dividing images into two components: reflection and illumination. Early Retinex-based methods mainly contain SSR [15], MSR [27] and MSRCR [14]. Guo [9] reconstruct the illumination map of the low-illumination image by finding the maximum value in the R, G, and B channels to estimate the ambient light. Li et al. [19] propose a noise map based on the Retinex model to enhance the robustness of the model. Ghosh et al. [8] propose a solution based on variational and filtering. Due to the nonlinearity between image channels and the complexity of data, it is difficult to accurately estimate the illumination component, which leads to excessive enhancement and local distortion in the above-mentioned methods.

2.2 Deep learning-based low-light enhancement approaches

Compared with traditional methods, deep learning based methods such as CNN can capture richer image features for low-light enhancement. Lore et al. [1, 22] propose an enhancement of shimmering images using a sparse denoising encoder, which suggests that deep learning can be used to construct shimmering enhancement models. Wang et al. [39] propose a multi-level low-light image enhancement method that decomposes the input images into two low-coupling feature components in the latent space. Guo et al. [10] propose a Zero-DCE, which learns curve function of the input images to achieve low-light enhancement performance. Recently, some researchers have also tried a variety of methods to achieve

[1] https://github.com/hlinqiao/MCLNet.git

low-light image enhancement, aiming to enhance the image details under complex dark backgrounds. Jiang et al. [13] design an EnlightenGAN for low-light enhancement using a non-reference strategy. Zhang et al. [51] propose a self-supervised low-light image enhancement method, which improves image contrast while reducing noise to avoid the blur in generated images. Fu et al. [7] design an adaptive low-light original image enhancement network. Zhang et al. [48] propose an effective Shuffle Attention (SA) module, which groups channels into multiple sub-features, and uses shuffle units for each sub-feature SA to depict the dependence of features in space and channel dimensions. Zhang et al. [47] proposed a new image reconstruction method that uses the surrounding environment information to reconstruct overexposed or oversaturated texture information for image detail reconstruction and applies edge-aware image decomposition for image enhancement. Syedet et al. [46] propose a multiscale residual module that contains key element, which extracts background information from multiple scales while retaining high-resolution spatial details.

2.3 Transformer-based low-light enhancement approaches

The Transformer is proposed by Vaswani et al. [34]. The Transformer's core is the self-attention mechanism that can be used to capture long-distance dependency between two pixels. Recently, the Transformer has been rapidly applied to computer vision tasks, such as object detection, image segmentation and super-resolution reconstruction. Dosovitskiy et al. [5] propose a ViT for image classification, which requires a lot of training data. Carion et al. [2] combine the CNN and the Transformer to propose an end-to-end detection framework DETR, which uses the CNN to extract basic image features and then sends it to the Transformer to capture long distance relations. Srinivas et al. [30] propose a BoTNet for object detection and instance segmentation by replacing ordinary convolution with the Transformer in the last three blocks of ResNet. Jiang et al. [12] combine the Transformer and the GAN to propose TransGAN for image generation. Wang et al. [40] propose a PVT by combining feature pyramid and the Transformer.

In this paper, we propose the MCLNet that inherits the advantages of CNN and the Transformer. Concretely, a sequence of the ARMBs is used to extract multiscale image features. And then, the CEM is designed to learn global information by using the Transformer. Such a mechanism allows our model to capture effective image information and long-range dependencies, thus improves low-light enhancement performance.

3 Methodology

In this section, we first introduce the overall framework of the proposed MCLNet. And then, we introduce the key components of MCLNet. Finally, we illustrate the loss function used.

3.1 Overall framework

The overall framework of the proposed MCLNet is shown in Fig. 1, which consists of two components: 1) an Effective Feature Extraction Subnetwork (EFES), 2) a Upsampling Subnetwork (US). Let I_{LL} and I_{EI} represent the low-light image and the enhanced image , respectively. The EFES consists of two parts: a Bottleblock of Scale Aggregation Module (BSAM) and a Context Encoding Module (CEM). Given a low-light image, the EFES first uses a convolution layer to obtain low-level feature. Then, to extract multiscale features, we

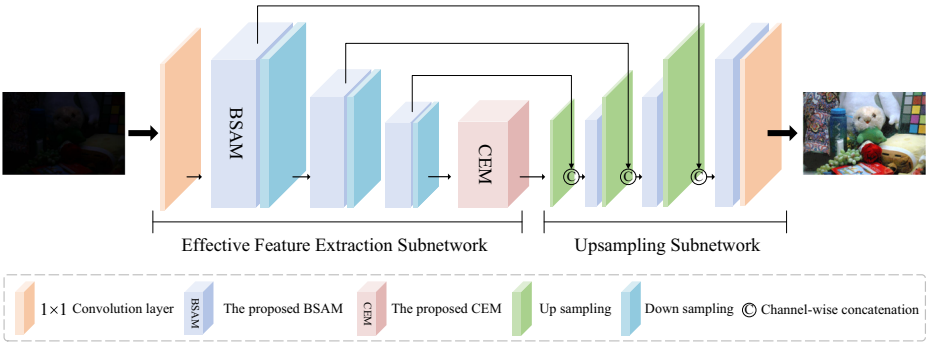

Fig. 1 Overview of Multiscale and Context Learning Network (MCLNet)

send the shallow information into BSAMs, and all the output of these intermediary modules are fed to the corresponding Upsampling Subnetwork (US) by skip connection. Finally, the output of the BSAMs are fed into the Context Encoding Module (CEM) for learning long-range relations. This stage can be formulated as:

$$F_{EFES} = H_{EFES}(I_{LL}) = H_{CEM}(H_{BD}(H_{LL}(I_{LL}))) \tag{1}$$

where H_{EFES} represents the EFES, which can be divided into the low-level information extraction $H_{LL}(\cdot)$, the multiscale learning H_{BD}, and the context feature modeling H_{CEM}. F_{EFES} means the output of the EFES.

Finally, the US processes the output feature map from the EFES by applying a convolution layer and skip connect to generate an enhanced image. These stages can be expressed as:

$$\begin{aligned} F_{US} &= H_{BSAM}(H_{UC}(F_{in}, F_{skip})) \\ I_{EI} &= H_{Conv}(F_{US}) \end{aligned} \tag{2}$$

where H_{BSAM} represents the BSAM. H_{UC} represents upsampling and concatenation operation. I_{EI} represents the perform of the convolution operation on F_{US}.

3.2 Effective feature extraction subnetwork (EFES)

The designed EFES is shown in Fig. 1, which is built by a convolution layer of three BSAMs and a CEM. It allows our method to focus on learning discriminative information at multiscale space and modeling long-range dependencies from high-level features, which is helpful for adaptive low-light enhancement.

Bottleblock of scale aggregation module (BSAM) To learn valuable hierarchical features, the proposed BSAM consists of Attentive Residual Multiscale Block (ARMB), Self-Calibrated Convolutions (SCConv), and ordinary convolutions (in Fig. 2). The BSAM extracts multiscale features by using the ARMB. Inspired by [35], we introduce the SCConv to adaptively build long-distance and inter-channel dependencies. In addition, we use Conv 1×1 to reduce the amount of calculation. The BSAM can be expressed as:

$$\begin{aligned} F_{multi} &= H_{ARMB}(w_1 x) \\ F_{BSAM} &= w_2 H_{Conv}(H_{SC}(F_{multi})) \end{aligned} \tag{3}$$

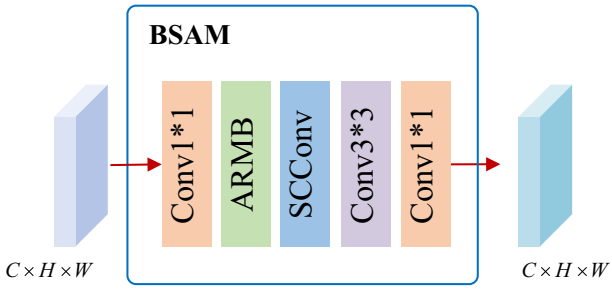

Fig. 2 The architecture of the Bottleblock of Scale Aggregation module (BSAM)

where $H_{ARMB}(\cdot)$, $H_{SC}(\cdot)$ and $H_{Conv}(\cdot)$ means the ARMB, the SCConv and the convolution layer. w_1 and w_2 are the weights of a dimension-reduction and a dimension-increasing layer.

Attentive residual multiscale block (ARMB) The position, shape and size of low-light regions are diverse and complex. Therefore, models with multiscale representation ability may improve low-light enhancement performance. In addition, different scale features have different noise information and semantic values. It is significant to focus on valuable features. We introduce the attention mechanism in multiscale representations learning to learn valuable multiscale information.

Multiscale Structure: Multiscale representations is important for most computer vision tasks. However, most previous CNN-based low-light enhancement methods have not considered both multiscale representations and filter noise information. Inspired by Inception [31] and PFANet [52], we design a multiscale module with spatial attention mechanism to extract valuable multiscale information.

As shown in Fig. 3, the ARMM consists of two branches. The first branch down samples the input features by two times to obtain 1/2 size features, the spatial attention mechanism [42] is used to filter noise information, 3×3 convolutional layers are used to extract the current scale features, and the upsampling operation is applied to restore the input features

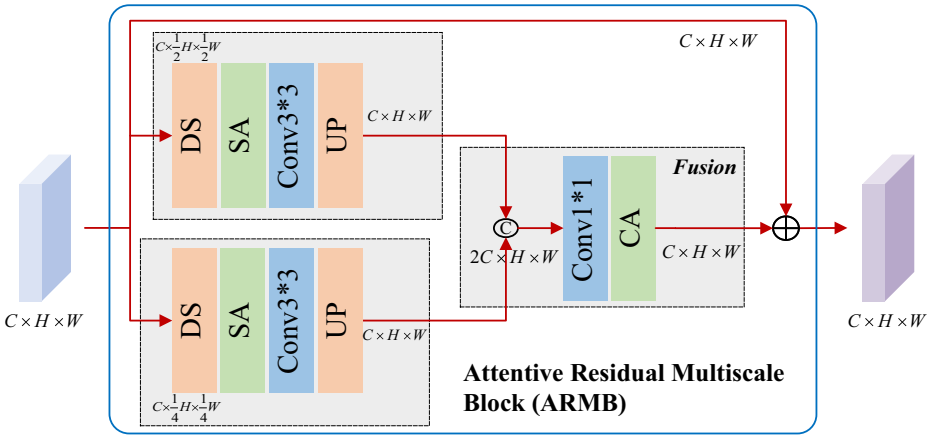

Fig. 3 The architecture of the Attentive Residual Multiscale Block (ARMB)

size. In the same way, the second branch down samples the input features by four times and obtains the features of the current scale through the same operation as the first branch.

Channel Attention Mechanism: Different scales of features may response to different low-light regions. We add channel attention [42] to weight different scales of features after multiscale feature fusion. The channel attention highlights the scale features which show a high response to low-light enhancement and help our method achieve adaptive enhancement ability. In addition, skip connections are used to directly add input features and multiscale features to make full use of shallow information.

Context encoding module (CEM) The CEM is performed by stacking the BSAMs and the Transformer in Fig. 4. The BSAMs in sequence is used to process deep semantic feature from the EFES and fuse the output of the Transformer. The CEM can be expressed as:

$$\begin{aligned} F_{CEM} &= H_{CEM}(x) \\ &= H_{BSAM}(H_{Trans}(H_{BSAM}(x))) \end{aligned} \tag{4}$$

where x is the input feature, $\mathrm{H}_{CEM}(\cdot)$, $\mathrm{H}_{BSAM}(\cdot)$, and $\mathrm{H}_{\mathrm{Trans}}(\cdot)$ are the CEM, BSAM, and Transformer, respectively.

Transformer: Due to the limitation of the size of the convolution kernel, it is difficult for CNN to obtain long-distance dependence in the image. In addition, low-illumination regions are often scattered. Therefore, Learning long-distance dependencies between regions facilitates the network to model the illumination distribution for adaptive enhancement performance. Therefore, the CEM introduces Transformer to capture effective long-range dependencies. The transformer is composed of a Patch Embedding, two Encoders, and a up-sampling operation.

Patch Embedding: The patch embedding converts the input feature into patches vectors. As shown in Fig. 4, the input feature ($C \times H \times W$) is first sent to a convolution layer and a normalization layer for obtaining patches vectors.

Encoder: The encoder is the core module of the Transformer. It is used to capture long-distance dependencies between pixels. Each encoder is composed of two Transformer Encoders. The input of Transformer Encoder is first normalized.Then linear transformation (w_1, w_2, w_3) is performed to obtain Key, Query and Value which are sent to the Multi-Head-Self-Attention (MHSA) (H_{MHSA}) to learn global relations. The Transformer Encoder

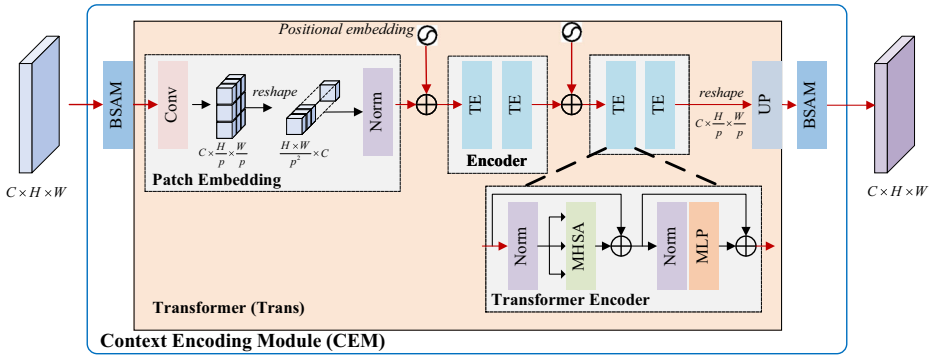

Fig. 4 The architecture of the Context Encoding Module (CEM)

can be expressed as:

$$\begin{aligned} F_{MHSA} &= H_{MHSA}(H_{Norm}(w_1x, w_2x, w_3x)) \\ F_{FF} &= H_{MLP}(H_{Norm}(F_{MHSA})) \oplus x \\ F_{TF} &= F_{FF} \oplus F_{MHSA} \end{aligned} \tag{5}$$

where x is input information. H, W are the height and width of the feature map. i, j are the element index of the feature map, and $\oplus$ is the pixel-by-pixel addition.

3.3 Upsampling subnetwork (US)

The EFES and CEM introduce the input low-illumination image above to extract effective features from the low-light images. The final high-quality enhancement results are obtained by stepwise reconstruction of the US. Concretely, the US consists of three sets of "Upsampling + concatenation + BSAM" and a Conv 1×1 stacking. As shown in Fig. 1, the features extracted by CEM are used as the initial input to US, and all intermediate inputs of EFES are concatenated with the upsampling results through a skip connection as the input to the corresponding BSAM in US, and Conv 1×1 is used to reduce the number of channels before the input to BSAM. In addition,the outputs of BSAMs are concatenated with the upsampling to enrich information. Finally, a Conv 1×1 is used to map the number of channels to the output channels to obtain the enhanced image.

3.4 Loss function

A joint loss function is designed by taking image reconstruction and content generation into account to guide the proposed model for generating high-quality enhanced images. The joint loss function can be formulated as:

$$L_{all} = w_0 L_{rec} + w_1 L_{per} \tag{6}$$

where L_{rec} and L_{per} mean the reconstruction loss and perceptual loss [16], respectively. We set $w_0 = 1$ and $w_1 = 0.006$.

The reconstruction loss is used to guide the MCLNet to generate enhanced images with complete structure by minimizing the error between the MCLNet output and the normal-illumination image. Let I_n and I_e denote the normal-illumination images and the enhanced images, respectively. The reconstruction loss can be expressed as:

$$L_{rec} = \frac{1}{2}(I_n - I_e)^2 \tag{7}$$

However, the reconstruction easily smooths the noise and details of the enhanced images simultaneously. Therefore, to preserve more texture information and improve quality of image content, the perceptual loss is introduced and formulated as:

$$L_{per} = \frac{1}{w_{ij} h_{ji} c_{ij}} \sum_{x=1}^{w_{ij}} \sum_{y=1}^{h_{ij}} \sum_{z=1}^{c_{ij}} \left(\emptyset_{ij} \left(I_l \right)_{xyz} - \emptyset_{ij} \left(I_n \right)_{xyz} \right)^2 \tag{8}$$

where w_{ij}, h_{ij}, c_{ij} are the size of each feature map in the MCLNet, $\emptyset_{ij}(.)$ represents the i-th convolution module of the j-th convolution group in the MCLNet, $\mu_{x,}\mu_y$ is the average pixel value, σ_x^2, σ_y^2 are variance, σ_{xy} is covariance, C_1, C_2 are constants.

4 Experiments

In this section, we first describe the experimental settings of our proposed method, including implementation details, datasets used, and evaluation metrics. Then, we compare the proposed method with state-of-the-art methods to show the efficiency of our network. Finally, we conduct a series of ablation studies to verify the effectiveness of the various components of our proposed network on low-light image enhancement tasks.

4.1 Implementation details

The inputs of MCLNet are a composite low-light image and a normal-light image. This paper randomly divides the input image into 256×256 image blocks during training. MCLNet uses ADAM optimizer ($\beta_1 = 0.9, \beta_2 = 0.99, \varepsilon = 10^{-8}$). The initial learning rate is $1e^{-3}$, each epoch attenuates 0.99, and a total of 8 epochs are trained. We use the Pytorch framework to implement MCLNet, and train on an NVIDIA Tesla P40. As shown in Fig. 5, we use gamma correction, gaussian noise and gaussian blur, respectively, to simulate low-illumination images from three aspects: illumination, noise and blur.

4.2 Datasets and evaluation metrics

To evaluate the performance of our proposed network, we conduct evaluations on three public datasets: LOL [41], NASA[2] and LIME [9]. LOL dataset [41] consists of 500 image pairs, where each pair contains a low-light image and its corresponding normal-light image. The first 485 image pairs are for training and the remaining are for testing. NASA is made up of 20 pairs of low-light images provided by NASA, including scenes of streets, houses and people, featuring uneven lighting and complex sources. LIME contains 10 low-light images.

To evaluate the model performance, we employ four popular image quality evaluation metrics for performance evaluation. (i) MultiScale-Structural Similarity Index (MS-SSIM) [36] is an index that measures the similarity of multiscale structure. MS-SSIM reflects the degree of image distortion, and the larger the value, the smaller the degree of image distortion. (ii) Peak Signal-to-Noise Ratio (PSNR) [37] is an objective measure of image distortion or noise level. The larger the PSNR value, the better the image quality and the closer to the original image. (iii) Natural Image Quality Evaluator (NIQE) [25] measures the degree of image distortion only using statistical rules observed in natural images and does not require human subjective scoring, and can measure the quality of images with arbitrary distortion. (iv) Blind Image Spatial Quality Evaluator (BRISQUE) [24] can measure the quality of images with common distortion such as compression artifacts, blurring, and noise.

4.3 Comparisons with state-of-the-art methods

MCLNet are compared with 13 excellent low-light image enhancement methods, including 9 deep learning methods (MSRNet [29], LLCNN [45], MBLLEN [33], GLAD [23], RetinexNet [41], KinD [49], EnlightenGAN [13], Zero-DCE [10], and Zero-DCE++ [3])

[2] https://dragon.larc.nasa.gov/retinex/pao/news/

Fig. 5 The process of synthetic a low-light image, which includes gamma correction, adding Gaussian noise and adding Gaussian blur

and 4 traditional methods (Dong [41], LIME [9], JED [4], Ying [28]). The above methods all use the author's open source code and parameter settings.

Quantitative evaluation The overall performance of the proposed model and other methods on the LOL dataset are shown in Table 1. We can see that the MCLNet achieves the best performance in terms of MS-SSIM, PSNR and NIQE, which proves that our method can obtain high-quality enhanced images. It demonstrates the effectiveness of MCLNet for low-light image enhancement tasks.

Qualitative evaluation Figure 6 shows the enhancement results of each method on the LOL dataset. Specifically, the enhancement results of Dong, JED, Ying and Zero-DEC are darker overall. The images enhanced by method MSRNet are distorted. The image colors enhanced by RetinexNet are oversaturated. MBLLEN, LLCNN and GLAD have improved

Table 1 Comparison of other methods and MCLNet on the LOL

Method	MS-SSIM ↑	PSNR ↑	NIQE ↓	BRSIQUE ↓
Dong [4]	0.626	14.995	18.864	98.889
LIME [9]	0.638	15.644	18.767	95.081
JED [28]	0.457	12.287	21.802	120.708
Ying [45]	0.644	17.203	19.466	108.096
MSRNet [29]	0.573	15.390	20.873	121.900
LLCNN [33]	0.691	16.140	19.685	108.283
MBLLEN [23]	0.735	16.599	20.495	107.795
GLAD [38]	0.723	17.345	20.287	99.887
RetinexNet [41]	0.712	16.774	18.733	85.788
KinD [49]	0.592	14.784	19.535	102.713
EnlightenGAN [13]	0.660	16.670	19.230	96.773
Zero-DCE [10]	0.585	14.860	19.459	108.127
Zero-DCE++ [3]	0.727	14.548	19.572	112.404
MCLNet (Ours)	**0.786**	**17.471**	**18.405**	**90.827**

Bold font indicates the top two methods. '↑' indicates the higher the better. '↓' means the lower the better

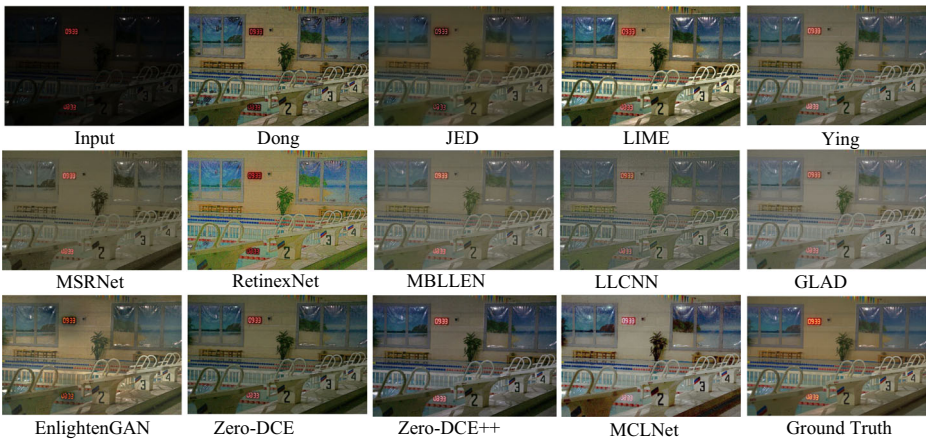

Fig. 6 Visual Comparison with state-of-the-art methods on the LOL [41] dataset

overall brightness but lower contrast, and EnlightenGAN have better color reproduction but are locally dark. The enhancement results of our proposed MCLNet method has achieved the best results in illuminance, color and contrast, significantly better than the other methods.

Figure 7 shows the comparison of enhancement results of each method on the LIME dataset. Specifically, the overall color of Dong, JED, LIME, EnlightenGAN, RetinexNet has color distortion and are reddish in color. MBLLEN and LLCNN have low contrast. In addition, Ying, MSRNet, and Zero-DEC have poor saturation. The enhancement results of our proposed MCLNet method have richer details, more natural colors and better visual effects. This is because ARMB learns the multiscale information of the low-illumination region and CEM captures the global contextual information.

Figure 8 shows the comparison of the enhancement results for each method on the NASA dataset. The environment of the input image belongs to foggy weather. Dong, JED, LIME,

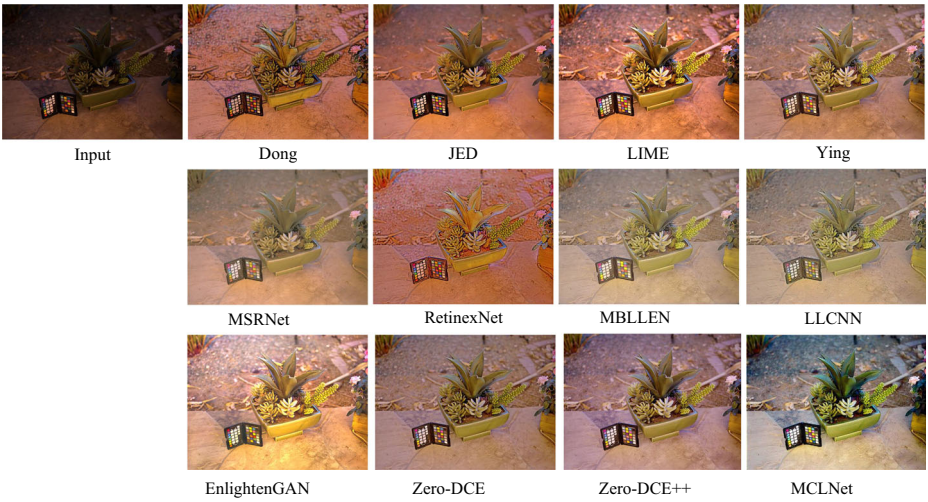

Fig. 7 Visual Comparison with state-of-art methods on the LIME [41] dataset

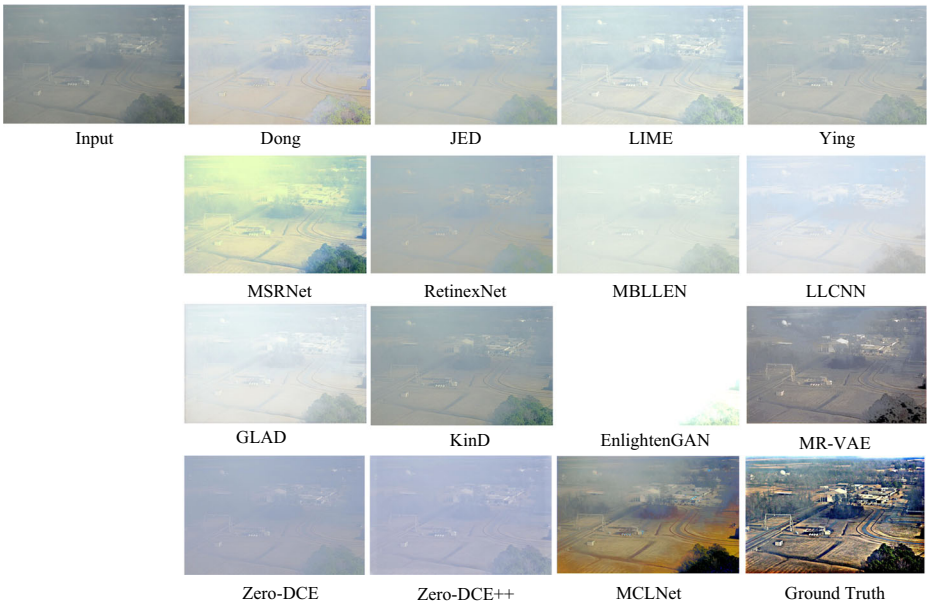

Fig. 8 Visual Comparison with state-of-the-art methods on the NASA[4] dataset

Ying, RetinexNet, MBLLEN and Zero-DCE basically play an enhancing effect, KinD can slightly enhance the image, and EnlightenGAN appears to be overexposed. Our method MCLNet can still effectively enhance the image in this type of environment, indicating that it has certain defogging ability and verifies that MCLNet can adaptively enhance the image.

In conclusion, the above experimental results show that our proposed MCLNet method can achieve superior performance compared to other methods.

4.4 Ablation study

The effectiveness of ARMB and CEM are evaluated by performing ablation experiments with MS-SSIM, PSNR, NIQE, and BRSIQUE. Note that the MCLNet is degraded to the baseline network in the ablation study by replacing the ARMB and CEM with ordinary convolution layers.

The effectiveness of ARMB and SCConv in the BSAM As shown in the 1-2 rows of Table 2, the proposed baseline's performance is improved by adding ARMB and SConv blocks,

Table 2 Effect of the ARMB, SCConv and CEM on the LOL in terms of MS-SSIM, PSNR, NIQE and BRSIQUE

Baseline	ARMB	SCConv	CEM	MS-SSIM ↑	PSNR ↑	NIQE ↓	BRSIQUE ↓
✓				0.707	16.989	21.347	128.277
✓	✓			0.758	17.308	20.014	100.434
✓	✓	✓		0.767	17.405	19.012	93.341
✓	✓	✓	✓	**0.786**	**17.471**	**18.405**	**90.827**

Fig. 9 Loss function ablation experiment of MCLNet on the LOL [41] dataset

which indicates that ARMB can extract multiscale features and SCConv can build long-distance and inter-channel dependencies.

The effectiveness of CEM The results of ablation study for CEM are shown in the fourth row of Table 2. The model's performance improve after adding CEM, demonstrating that the CEM can capture long-distance dependencies. It allows the model to be adaptively enhanced to various regions of illumination.

The effectiveness of loss function The evaluation of loss function is shown in Fig. 9. It can be seen that the contrast and saturation of enhanced results are poor when using only L_{rec}. Adding the L_{per} makes the details clearer,and the visual effects are better in enhanced outputs. These experimental results show that the introduction of L_{per} loss can effectively promote our model to generate higher-quality images.

5 Conclusion

This paper proposes MCLNet for low-light image enhancement, which is composed of Efficient Feature Extraction Subnetwork (EFES) and Upsampling Subnetwork (US). Concretely, the EFES consists of a Bottleblock of Scale Aggregation Module (BSAM) and a Context Encoding Module (CEM). The core part of BSAM is an Attentive Residual Multiscale Block (ARMB), which can extract multiscale features. The CEM captures the context information of entire image. The Transformer in CEM captures the long distance dependencies between local regions of images. These designs can not only enhance more detailed information of low-light images, but also avoid over or under enhancement. The experimental results on three public datasets demonstrate that our proposed method outperforms the other state-of-the-art methods. In the future, we will further explore Transformer and low-light image enhancement methods focusing on detail enhancement and adaptability.

Funding This work was supported in part by the National Natural Science Foundation of China under grant 62072169 and 62172156, and Natural Science Foundation, and Natural Key R&D Program of China under Grant No.2020YFB1713003,and Scientific Research Project of Hunan Provincial Education Department No.19A286.

Data Availability The dataset analyzed in the current study is available from the corresponding author upon reasonable request. The models and code analyzed in the current study will soon be publicly available at https://github.com/hlinqiao/MCLNet.git.

Declarations

Conflict of Interests The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cai Z, Zhang Y, Manzi M, Oztireli C, Gross M, Aydin TO (2021) Robust image denoising using kernel predicting networks
2. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision (ECCV). Springer, pp 213–229
3. Chongyi L, Guo C, Loy CC (2021) Learning to enhance low-light image via zero-reference deep curve estimation. arXiv:2103.00860
4. Dong X, Wang G, Pang Y, Li W, Wen J, Meng W, Lu Y (2011) Fast efficient algorithm for enhancement of low lighting video. In: 2011 IEEE international conference on multimedia and expo (ICME), pp 1–6
5. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929
6. Fu X, Zeng D, Huang Y, Zhang X-P, Ding X (2016) A weighted variational model for simultaneous reflectance and illumination estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2782–2790
7. Fu Q, Di X, Zhang Y (2020) Learning an adaptive model for extreme low-light raw image processing. arXiv:2004.10447
8. Ghosh S, Chaudhury KN (2019) Fast bright-pass bilateral filtering for low-light enhancement. In: 2019 IEEE international conference on image processing (ICIP). IEEE, pp 205–209
9. Guo X, Li Y, Ling H (2016) Lime: low-light image enhancement via illumination map estimation. IEEE Trans Image Process 26(2):982–993
10. Guo C, Li C, Guo J, Loy CC, Hou J, Kwong S, Cong R (2020) Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1780–1789
11. Huang S-C, Cheng F-C, Chiu Y-S (2012) Efficient contrast enhancement using adaptive gamma correction with weighting distribution. IEEE Trans on Image Process 22(3):1032–1041
12. Jiang Y, Chang S, Wang Z (2021) Transgan: two transformers can make one strong gan. 1(3) arXiv:2102.07074
13. Jiang Y, Gong X, Liu D, Cheng Y, Fang C, Shen X, Yang J, Zhou P, Wang Z (2021) Enlightengan: deep light enhancement without paired supervision. IEEE Trans Image Process 30:2340–2349
14. Jobson DJ, Rahman Z-U, Woodell GA (1997) A multiscale retinex for bridging the gap between color images and the human observation of scenes. IEEE Trans on Image Process 6(7):965–976
15. Jobson DJ, Rahman Z-U, Woodell GA (1997) Properties and performance of a center/surround retinex. IEEE Trans Image Process 6(3):451–462
16. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision (ECCV). Springer, pp 694–711
17. Land EH (1964) The retinex. Am Sci 52(2):247–264
18. Li J, Fang F, Mei K, Zhang G (2018) Multi-scale residual network for image super-resolution. In: Proceedings of the european conference on computer vision (ECCV), pp 517–532
19. Li M, Liu J, Yang W, Sun X, Guo Z (2018) Structure-revealing low-light image enhancement via robust retinex model. IEEE Trans Image Process 27(6):2828–2841
20. Li X, Guo X, Mei L, Shang M, Gao J, Shu M, Wang X (2020) Visual perception model for rapid and adaptive low-light image enhancement. arXiv:2005.07343
21. Liu C, Sui X, Liu Y, Kuang X, Li G, Chen Q (2019) Adaptive contrast enhancement based on histogram modification framework. J Mod Opt 66(15):1590–1601
22. Lore KG, Akintayo A, Sarkar S (2017) Llnet: a deep autoencoder approach to natural low-light image enhancement. Pattern Recogn 61:650–662

23. Lv F, Lu F, Wu J, Lim C (2018) Mbllen: low-light image/video enhancement using cnns. In: British machine vision association (BMVC), p 220
24. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Tran on Image Process 21(12):4695–4708
25. Mittal A, Soundararajan R, Bovik AC (2012) Making a "completely blind" image quality analyzer. IEEE Sig Process Lett 20(3):209–212
26. Mun J, Jang Y, Nam Y, Kim J (2019) Edge-enhancing bi-histogram equalisation using guided image filter. J Vis Commun Image Represent 58:688–700
27. Rahman Z-U, Jobson DJ, Woodell GA (1996) Multi-scale retinex for color image enhancement. In: Proceedings of 3rd IEEE international conference on image processing, vol 3. IEEE, pp 1003–1006
28. Ren X, Li M, Cheng W-H, Liu J (2018) Joint enhancement and denoising method via sequential decomposition. In: 2018 IEEE international symposium on circuits and systems (ISCAS). IEEE, pp 1–5
29. Shen L, Yue Z, Feng F, Chen Q, Liu S, Ma J (2017) Msr-net: low-light image enhancement using deep convolutional network. arXiv:1711.02488
30. Srinivas A, Lin T-Y, Parmar N, Shlens J, Abbeel P, Vaswani A (2021) Bottleneck transformers for visual recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 16519–16529
31. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI conference on artificial intelligence
32. Tan SF, Isa NAM (2019) Exposure based multi-histogram equalization contrast enhancement for non-uniform illumination images. IEEE Access 7:70842–70861
33. Tao L, Zhu C, Xiang G, Li Y, Jia H, Xie X (2017) Llcnn: a convolutional neural network for low-light image enhancement. In: 2017 IEEE visual communications and image processing (VCIP). IEEE, pp 1–4
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
35. Wang Y, Chen Q, Zhang B (1999) Image enhancement based on equal area dualistic sub-image histogram equalization method. IEEE Trans Consum Electron 45(1):68–75
36. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. In: The thrity-seventh asilomar conference on signals, systems & computers, 2003, vol 2. IEEE, pp 1398–1402
37. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
38. Wang W, Wei C, Yang W, Liu J (2018) Gladnet: low-light enhancement network with global awareness. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 751–755
39. Wang L, Fu G, Jiang Z, Ju G, Men A (2019) Low-light image enhancement with attention and multi-level feature fusion. In: 2019 IEEE international conference on multimedia & expo workshops (ICMEW). IEEE, pp 276–281
40. Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L (2021) Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. arXiv:2102.12122
41. Wei C, Wang W, Yang W, Liu J (2018) Deep retinex decomposition for low-light enhancement. arXiv:1808.04560
42. Woo S, Park J, Lee J-Y, Kweon IS (2018) Cbam: convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19
43. Xu K, Yang X, Yin B, Lau RWH (2020) Learning to restore low-light images via decomposition-and-enhancement. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2281–2290
44. Yang C, Qiao S, Kortylewski A, Yuille A (2021) Locally enhanced self-attention: rethinking self-attention as local and context terms. arXiv:2107.05637
45. Ying Z, Li G, Ren Y, Wang R, Wang W (2017) A new low-light image enhancement algorithm using camera response model. In: Proceedings of the IEEE international conference on computer vision workshops, pp 3015–3022
46. Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H, Shao L (2020) Learning enriched features for real image restoration and enhancement. In: European conference on computer vision (ECCV) 2020: 16th, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Springer, pp 492–51
47. Zhang Y, Aydın TO (2021) Deep hdr estimation with generative detail reconstruction. In: Computer graphics forum, pp 179–190
48. Zhang Q-L, Yang Y-B (2021) Sa-net: shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2235–2239

49. Zhang Y, Zhang J, Guo X (2019) Kindling the darkness: a practical low-light image enhancer. In: Proceedings of the 27th ACM international conference on multimedia, pp 1632–1640
50. Zhang C, Yan Q, Zhu Yu, Li X, Sun J, Zhang Y (2020) Attention-based network for low-light image enhancement. In: IEEE international conference on multimedia and expo (ICME). IEEE, pp 1–6
51. Zhang Y, Di X, Zhang B, Li Q, Yan S, Wang C (2021) Self-supervised low light image enhancement and denoising. arXiv:2103.00832
52. Zhao T, Wu X (2019) Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
53. Zhuang L, Guan Y (2018) Adaptive image enhancement using entropy-based subhistogram equalization. Comput Intell Neurosci 2018