MARN: Multi-Scale Attention Retinex Network for Low-Light Image Enhancement

XIN ZHANG AND XIA WANG

MOE Key Laboratory of Photo-Electronic Imaging Technology and Systems, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China

Corresponding author: Xia Wang (angelniuniu@bit.edu.cn)

ABSTRACT

Images captured in low-light conditions often suffer from bad visibility, e.g., low contrast, lost details, and color distortion, and image enhancement methods can be used to improve the image quality. Previous methods have generally obtained a smooth illumination map to enhance the image but have ignored details, leading to inaccurate illumination estimations. To solve this problem, we propose a multi-scale attention retinex network (MARN) for low-light image enhancement, which learns an image-to-illumination mapping to obtain a detailed inverse illumination map inspired by retinex theory. In order to introduce more image priors, we introduce a novel illuminance-attention map to guide the model to characterize varying-lighting areas, which we combine with the low-light image as the model input. MARN consists of a multi-scale attention module and a feature fusion module; the former extracts multi-resolution features with attention-based feature aggregation, while the latter further merges the output features of the previous module with the input. To achieve better visibility, we formulate a novel loss function to synthetically measure the illumination, detail, and colorfulness of the image. Extensive experiments are performed on several benchmark datasets. The results demonstrate that our method outperforms other state-of-the-art methods according to both objective and subjective metrics.

INDEX TERMS

Low-light image enhancement, deep learning, multi-scale attention retinex network.

I. INTRODUCTION

Because of environmental or technical constraints, images are often captured under complicated lighting conditions. Images captured in low-light environments suffer from poor image quality in terms of low visibility, low contrast, and color distortion. These images not only look unpleasant but also make post-processing such as object detection and segmentation, more difficult. Therefore, it is necessary to determine how to address these problems.

Many approaches have been proposed for low-light image enhancement. Traditional image enhancement methods mainly include histogram equalization [1]–[3] and retinex theory-based methods [4]–[6]. The histogram equalization method maintains the relative relationships between the pixel values, so that the pixel values are uniformly distributed, which contributes to the extension of the image histogram. Retinex theory is effective for image enhancement. Retinex-based methods demonstrate that images can be decomposed into reflection and illumination components. Reflection is an inherent property of a scene, while illumination is affected by the environmental illuminance. Retinex-based methods usually enhance the illumination component of a low-light image to approximate the corresponding normal-light image.

In recent years, deep learning methods have been widely used in computer vision tasks and have achieved significant breakthroughs in multiple fields, such as image denoising [7]–[9], image super-resolution [10], [11], object detection [12], [13], and image enhancement [14]–[17], [19]. Convolutional neural networks (CNNs) learn strong priors from large-scale datasets. Low-light image enhancement can be thought of as an image restoration task. Wang et al. [43] reviews the main techniques of low-light image enhancement developed over the past decades. Some methods directly learn end-to-end mapping from a low-light image to a normal-light image (e.g., GLADNet [14] and LLCNN [15]), which is a general application of CNN. Some methods consist of several CNNs that learn the illumination and the reflectance separately (e.g., Retinex-Net [16] and KinD [17]).
However, ground truth images of illumination and reflectance are difficult to obtain and the fusion of multiple networks greatly increases the training difficulty. Some methods are unsupervised and apply the structure of generative adversarial networks (GANs) [18] or non-reference loss functions to estimate normal-light images from low-light images. GANs consist of a generator and a discriminator; however, it is difficult to train two networks simultaneously. Zero-DCE [19] reformulates the task as an image-specific curve estimation problem and only uses several non-reference loss functions. However, the desired enhanced result cannot be easily obtained when only relying on non-reference loss functions.

To predict a well-detailed illumination map for low-light image enhancement, in this paper, we propose a novel framework called multi-scale attention retinex network (MARN). First, in order to introduce more image priors, our method transforms the image into an illuminance-attention map and combines this map with the input image as the network input. Based on retinex theory, MARN is designed to predict a detailed inverse illumination map that contains the detail and color information. As opposed to multiple proposed CNN methods, we propose a novel multi-scale attention module for feature extraction, which can greatly improve the generalization capability of the network. Further, we adopt a loss function that comprehensively covers the structure similarity, details, and color restoration.

Our contributions are summarized as follows:

- We propose a novel method, MARN, to obtain a detailed inverse illumination map for low-light image enhancement. In order to introduce more image priors, we introduce a novel illuminance-attention map to guide the model to characterize varying-lighting areas, then we combine the low-light image with the illuminance-attention map as the network input which guides the network to estimate a more accurate illumination distribution.
- Both global and local features are necessary for low-light image enhancement. We design a novel and effective multi-scale attention module to obtain the features at different scales and integrate them at the main resolution using attention-based feature aggregation.
- We introduce a novel and comprehensive loss function for the illumination, details, and color restoration. The loss function contains reference and non-reference components that improve the visibility of the results.
- Comprehensive experiments are conducted on several benchmark datasets and demonstrate that our method outperforms existing state-of-the-art methods. Additionally, an ablation study is conducted to demonstrate the efficacy of our structure.

II. RELATED WORK

**Traditional methods** Histogram equalization-based methods maintain the relative relationships between the pixel values, so that the pixel values are uniformly distributed, which contributes to the extension of the image histogram. Particularly, dynamic histogram equalization [3] divides the histogram into several parts and performs histogram equalization processing in each sub-hierarchy. Contrast limit adaptive histogram equalization [2] can adaptively limit the range of the enhancement effects after histogram equalization. However, this may result in over- and under-enhancement. Other methods based on retinex theory often decompose the image into illumination and reflectance. Early attempts, e.g., the MSR [5] and SSR [6] methods, committed to recovering the illumination map for low-light image enhancement. Recently, the LIME method [20] proposed a refined illumination map to estimate the enhancement. The NPE method [21] balances the details and naturalness of images. The SRIE method [22] designs a weighted variational model to simultaneously estimate the reflectance and the illumination. The LR3M method [44] inject low-rank prior to estimate a piece-wise smoothed illumination and a noise-suppressed reflectance. While the RIO-NS method [45], [46] proposed a regularized illumination optimization method to enhance maritime images. However, most retinex-based methods may cause serious color distortion and cannot properly enhance images with relatively high dynamic ranges.

**Learning-based methods** In recent years, deep learning methods have been widely used in low-level computer vision tasks. LLNet [23] was one of the earliest methods to implement a stacked auto-encoder for low-light image enhancement and denoising. LLCNN [15] implemented a CNN-based module to perform low-light image enhancement. SID [24] is designed to deal with raw short-exposure low-light images, while GLADNet [14] consists of two sub-networks, a global illumination estimation step, and a detail reconstruction step. MBLENN [25] uses a novel multi-branch network structure to learn mappings from low-light images to normal-light images. Retinex-Net [16] is based on retinex theory and enhances the illumination part while denoising the reflection part, the final enhanced image is obtained after fusion. Similarly, KinD [17] designs a retinex-based network by adding a restoration-net for noise removal. DeepUPE [28] learns an image-to-illumination mapping to predict a smooth illumination map. Most of these methods are trained on synthesized paired datasets using gamma correction, which may cause significant color distortion.

Unsupervised learning has also been used for low-light image enhancement. EnlightenGAN [26] proposes a highly effective unsupervised GAN for low-light enhancement. Zero-DCE [19] formulates light enhancement as an image-specific curve estimation task with a deep network and does not require any paired or unpaired data when training. However, satisfactory results cannot be achieved without paired supervision.

Our method incorporates the advantages of previous methods by applying the illuminance-attention map as a guidance to estimate a detailed inverse illumination map instead of a smooth one. Additionally, we employ the MIT-Adobe
FiveK [27] dataset as a training set and combine reference loss functions and non-reference loss functions for better visibility. The proposed multi-scale attention module can effectively improve the generalization capabilities and learning abilities of networks.

III. THE PROPOSED METHOD

In this section, we introduce the proposed MARN method, as illustrated in Fig. 1. We then provide the details of the network.

A. RETINEX-BASED ENHANCEMENT MODEL

As previously mentioned, retinex theory has been widely used for low-light image enhancement, which can be formulated as

\[ S(x, y) = R(x, y) \ast I(x, y) , \]  

where \((x, y)\) represents the position of each pixel, \(\ast\) represents pixel-wise multiplication, \(S\) represents the observed image, \(R\) represents the reflectance component, and \(I\) represents the illumination component. Similar to other retinex-based methods, we treat the reflectance \(R\) as the enhanced result and \(S\) as the input low-light image. Therefore, once \(S\) is obtained, the enhanced result \(R\) is calculated as \(R = S/I\).

Existing methods, such as DeepUPE [28] and ReANet [29], use CNNs to estimate \(I\). However, the division operator may cause problems, e.g., gradient explosion. In our method, we estimate the inverse illumination component \(I^{-1}\) directly and the final result can be calculated as \(R = S \ast I^{-1}\). The inverse illumination component \(I^{-1}\) has three channels (\(R, G, B\)) to match the non-uniformity of the different channels.

B. THE ILLUMINANCE-ATTENTION MAP

Instead of directly employing the low-light image as the input, we introduce the illuminance-attention map, which is different from the inverse illumination map \(I^{-1}\). Our motivation is to provide more prior information for the network to allow the network to pay more attention to the underexposed areas and to avoid over-enhancing the normally exposed areas. The illuminance-attention map is formulated such that

\[ IA = \min_c \left( \frac{\max_c (S) - S}{\max_c (S) - \min_c (S)} \right) \]  

where \(\max_c(\cdot)\) and \(\min_c(\cdot)\) represent the maximum and minimum values of three color channels, respectively, and \(S\) is the input low-light image.

As shown in Fig. 2(b), our illuminance-attention map indicates the underexposed areas. Higher illumination results in lower illuminance-attention map values. Meanwhile, as shown in Figs. 2(c)–2(e), our inverted illuminance-attention map is similar to the illumination map introduced by several retinex-based methods, such as LIME and SRIE. This infers that the prior information brought by our illuminance-attention map is meaningful. Additionally, our map provides more details than other maps. Therefore, we employ the illuminance-attention map as guidance for our model.

C. NETWORK STRUCTURE

The MARN pipeline is illustrated in Fig. 1. MARN first employs the low-light image \(S\) and the illuminance-attention map \((IA)\) as the input, then it goes through the multi-scale attention module and feature fusion module to predict the inverse illumination map such that

\[ I^{-1} = \mathcal{F}(S, IA) , \]  

where \(\mathcal{F}(\cdot)\) represents the entire pipeline of the network, and the enhanced result \(R\) is obtained such that

\[ R = S \ast I^{-1} . \]
(2) the resolution of feature maps decreases gradually from high to low and then increases, such as Unet [14], [16], [17]; and (3) dilated convolutions are applied to enlarge the receptive field [32]. Such designs either ignore or lose features of different resolutions during the convolution. We designed a novel CNN structure called the multi-scale attention module, which is able to maintain high resolution during the entire process and transform the low-resolution features to high resolution at every step.

In this paper, the multi-scale attention module is also able to fuse global and local features by fusing high- and low-resolution features because a smaller resolution corresponds to a larger receptive field. Global features, represented by low resolution, consist of the average illumination, color distribution, and scene category, while local features, represented by high resolution, consist of the details, contrast, and noise. The multi-scale attention module contains a series of multi-scale attention blocks (MABs), which are capable of producing several spatially precise multi-channel outputs, as shown in Fig. 3(a). The MABs maintain the high-resolution representations by aggregating the rich contextual information from the low-resolution branches. Each MAB consists of multiple branches in parallel (three branches in our method). The stream that has the same resolution as the input is the main branch, while the others are subordinate branches. The details are shown below.

As illustrated in Fig. 3, each MAB adopts three inputs \( (x_{s1}, x_{s2}, x_{s3}) \) and produces three outputs \( (y_{s1}, y_{s2}, y_{s3}) \) simultaneously, where \( s_1 = 1, s_2 = 1/2, \) and \( s_3 = 1/4. \) Because each MAB requires three different resolution feature input maps, we transform the input \( (S, IA) \) into three scales using bilinear down-sampling for the first MAB. Then, each subsequent block applies the output of the previous block as input. The different scaled inputs are sent to different branches:

\[
\begin{align*}
fs_1 &= F_{s1}(x_{s1}), \\
fs_2 &= F_{s2}(x_{s2}), \\
fs_3 &= F_{s3}(x_{s3}).
\end{align*}
\]  

(5)

where \( F_{si} (i = 1, 2, 3) \) represents the feature extraction at the different branches, which employ several residual blocks [36]. The structure of a residual block is introduced in Fig. 3(c). Then, \( fs_2 \) and \( fs_3 \) are up-sampled via convolution layers and merged with \( fs_1 \):

\[
fs_{merge} = (fs_1, F_{u×2}(fs_2), F_{u×4}(fs_3))
\]  

(6)

Here, \( F_{u×2}() \) and \( F_{u×4}() \) indicate 2× and 4× up-sampling, respectively.

The multi-scale information is fused across the different branches, and we employ an attention block to filter the fused feature map to improve the feature extraction. Inspired by several attention-based computer vision tasks [39]–[41], we introduce a dual attention block to filter the features from the spatial and channel dimensions. The structure of a dual attention block is shown in Fig. 3(b). Given an input, \( w_{s1} \in \mathbb{R}^{H \times W \times C} \), the dual attention block contains two streams processing \( w_{s1} \). The channel attention stream adopts global average pooling and global max pooling across the spatial dimension to generate the channel attention operators \( C_{avg} \in \mathbb{R}^{1 \times 1 \times C} \) and \( C_{max} \in \mathbb{R}^{1 \times 1 \times C} \), and the output of the channel attention stream \( F_C \in \mathbb{R}^{H \times W \times C} \) is obtained by summing \( C_{avg} \) and \( C_{max} \). The spatial attention stream aims to generate a spatial-wise attention map. Similar to the channel attention stream, it applies global average pooling and max pooling across the channel dimension to generate the spatial attention operators \( S_{avg} \in \mathbb{R}^{H \times W \times 1} \) and \( S_{max} \in \mathbb{R}^{H \times W \times 1} \), then...
$S_{avg}$ and $S_{max}$ pass through a convolution layer to obtain the output spatial attention map $F_S \in \mathbb{R}^{H \times W \times 1}$, $F_C \in \mathbb{R}^{1 \times 1 \times C}$ and $F_S \in \mathbb{R}^{H \times W \times 1}$ are used to rescale $w_s\gamma$.

The output of the multi-scale attention block is

$$y_3 = F_C \left(f_{merge}\right) \times F_S \left(f_{merge}\right), y_2 = f_{s_2}, y_3 = f_{s_3}. \quad (7)$$

2) FEATURE FUSION MODULE

The multi-scale attention module extracts a large number of features from the input, and its output cannot be used as the final output of the network. The motivation for this feature fusion module is to better integrate the useful information extracted from the preceding multi-scale attention module. To better preserve the details and textures of the image, we introduce the input image again at this stage. This module contains three residual blocks. The output of the feature fusion module is the inverse illumination map.

$$I^{-1} = F_D \left(M \left(x\right), x\right), \quad (8)$$

where $M \left(x\right)$ is the output of the multi-scale attention module, $x$ is the input tensor, and $F_D()$ represents the feature fusion module. Despite the simple appearance of the feature fusion module, it plays an important role in our network.

D. LOSS FUNCTIONS

To achieve better visibility, we applied four loss functions to measure the output of our network, i.e.,

$$L_{total} = \lambda_r L_r + \lambda_s L_{ssim} + \lambda_c L_{color} + \lambda_d L_{detail}. \quad (9)$$

where $L_r$, $L_{ssim}$, $L_{color}$, and $L_{detail}$ represent the reconstruction loss, structure similarity loss, color loss, and detail loss, respectively, and $\lambda_r$, $\lambda_s$, $\lambda_c$, and $\lambda_d$ are their corresponding coefficients. In order to better balance the various loss functions, we successively introduce different loss functions in the experiment, so as to obtain a satisfactory coefficients corresponding to each loss function. After several experiments, we set their values to $\lambda_r = 0.8$, $\lambda_s = 0.6$, $\lambda_c = 0.1$, and $\lambda_d = 0.3$. The details of the four loss functions are given below.

1) RECONSTRUCTION LOSS

In this paper, we apply the robust Charbonnier $L_1$ loss function [33] to measure the difference between the predicted image and the ground truth. Compared to the $L_1$ loss function, it can better handle outliers and improve the model performance. It is expressed as

$$L_r = \sqrt{\left\|R_{GT} - S \ast I^{-1}\right\|^2 + \varepsilon} \quad \left(\varepsilon = 10^{-6}\right), \quad (10)$$

where $R_{GT}$ is the ground truth image, $S$ is the input low-light image, and $I^{-1}$ is the output inverse illumination map.

2) STRUCTURE SIMILARITY LOSS

Low-light images always suffer from structural distortion problems such as image blurring. To better preserve the structural information of an image, we applied the structure similarity (SSIM) [34] as one of our loss functions. The SSIM value ranges from 0 to 1, where higher values represent better similarity.

The SSIM value for a pixel $p$ is computed such that

$$SSIM \left(x, y\right) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}. \quad (11)$$

Accordingly, the SSIM loss function can be computed as

$$L_{ssim} = 1 - \text{SSIM} \left(R_{GT}, S \ast I^{-1}\right). \quad (12)$$

3) COLOR LOSS

Color correction issues are often ignored in low-light enhancement tasks. Despite the use of paired images in training, different datasets have different color styles. To improve the natural color fidelity of enhanced images, we introduce a novel non-reference color loss function. Reference [35] has shown that the image colorfulness can be represented effectively by the CIQI colorfulness metric. The non-reference CIQI metric is formulated as

$$\text{CIQI} = \frac{\sigma_a^2 + \sigma_b^2 + 0.3\sqrt{\mu_a^2 + \mu_b^2}}{85.59}, \quad (13)$$

where $\alpha = R - G$ and $\beta = 0.5 \left(R + G\right) - B$ are opponent red–green and yellow–blue spaces, respectively, and $\sigma_a^2$ and $\sigma_b^2$ and $\mu_a$ and $\mu_b$ represent the variance and mean values along these two opponent color axes, respectively. Higher CIQI scores indicate better natural color fidelity. Inspired by the CIQI metric, we designed a non-reference color loss function, which is formulated such that

$$L_{color} = 1 - 0.5 \times \sqrt{\mu_a^2 + \mu_b^2}. \quad (14)$$

where $\mu_a$ and $\mu_b$ are the mean values of the enhanced result.

4) DETAIL LOSS

Image gradients are generally used for edge detection and noise detection. In the gradient calculation, first-order (Sobel and/or Prewitt) and second-order (Laplacian) derivative operators are widely employed. Compared to the first-order derivative, the second-order derivative has a stronger response to fine details and is more sensitive to noise. Therefore, in this paper, we designed a Laplacian-based gradient loss function to restore lost details in the low-light images and to smooth the noise simultaneously. The detail loss function is formulated such that

$$L_{detail} = \left\|\text{Grad} \left(R_{GT}\right) - \text{Grad} \left(S \ast I^{-1}\right)\right\|, \quad (15)$$

where Grad() represents the Laplacian operator.

IV. EXPERIMENTAL SETTING

Dataset Multiple previous studies have generated synthesized low-light images for training via gamma correction. However, a gamma correction is a global adjustment for an entire image and may cause color distortion, which is not suitable for low-light image enhancement. We employed the
MIT-Adobe FiveK dataset [27] as our training set. It consists of 5000 raw images. Each of these images was retouched by different experts (A/B/C/D/E) using Adobe Lightroom software. We used the output by Expert E as our ground truth. We randomly selected 4000 images for training, and the remaining 1000 images were used for validation and testing. All the images were resized to 400 × 600 and converted to Portable Network Graphics format.

Implementation details We implemented our model using the PyTorch framework. The Adam optimization algorithm was used, with \( \beta_1 = 0.9 \) and \( \beta_2 = 0.999 \). The initial value of the learning rate was \( 1 \times 10^{-3} \), and the decaying coefficient was 0.5 every 150 epochs. All the experiments were performed on two NVIDIA GeForce GTX1080TI GPUs.

A. ABSTRACTION STUDY

In this section, we describe several ablation experiments we performed to prove the effectiveness of each component of our model.

1) EFFECT OF THE ILLUMINANCE-ATTENTION MAP

As previously mentioned, we combined the illuminance-attention (IA) map with the low-light image to form the network input. To evaluate the effect of the IA map, we also trained a model without it. The results in Table 1 show that the IA map greatly improves the model performance, where \( \Delta E \) represents the color difference.

2) EFFECT OF THE MULTI-SCALE ATTENTION MODULE

To verify the effect of our multi-scale attention module, we made a comparison with ResNet, which consists of residual blocks stacked one by one. The number of residual blocks was the same as that in our multi-scale attention module. As shown in Table 1, compared to the residual blocks, our multi-scale attention module greatly improves the generalization capability and the performance of the network.

3) EFFECT OF THE LOSS FUNCTIONS

We also verified the effects of the different loss functions in our model: L1 loss, SSIM loss, detail loss, and color loss. As shown in Table 1, the results demonstrate the importance of each loss function and that all of the loss functions are indispensable in our method.

B. EVALUATION ON BENCHMARK DATASETS

We compared MARN to several state-of-the-art low-light enhancement methods, including SRIE [22], Retinex-Net [16], MBLLEN [25], KinD [17], and Zero-DCE [19]. The results were reproduced using the publicly available source codes provided by the authors. We performed experiments on several benchmark datasets, including the DICM [37], Fusion, LIME [20], MEF [38], NPE [21], LOL, and MIT-Adobe FiveK datasets. Of these, LOL and MIT-Adobe FiveK consist of paired images, while the others only contain low-light images. LOL is widely used in low-light image enhancement and includes 500 low/normal-light image pairs; 485 pairs were used for training, while 15 were used for testing. MIT-Adobe FiveK is the training set used for our method, and we selected the last 30 images for testing from the remaining 1000 images.

1) VISUAL COMPARISON

Figs. 4–7 show a visual comparison of these methods. We can see from the images that SRIE and MBLLEN do not sufficiently improve the image illumination; their results appear dim compared to the others. Retinex-Net and KinD greatly improve the image brightness; however, they also cause significant color distortion, which makes the images look unnatural, especially Retinex-Net. In Fig. 6, we see that KinD enhances the images unevenly. Zero-DCE provides a fairly good result when enhancing images with higher brightness, as shown in Fig. 7; however, it does not work when dealing with images with low overall brightness, such as the trees in Fig. 6. Our method better improves the brightness and details in all areas of the image and prevents the occurrence of over/underexposure across all the benchmark datasets. As shown in Fig. 4 and Fig. 5, our method can better remove the random noise in the images. Additionally, our method better restores the colorfulness of the image itself, which makes the enhanced effect look more natural and the colors more vivid, such as the tower and the woman in Fig. 7. In Fig. 6, our result looks even better than the ground truth image.

2) QUANTITATIVE COMPARISON

To verify the learning effectiveness and enhancement capability of our model, we compared it to other state-of-the-art methods using several objective evaluation metrics. Because the datasets used for comparison can be divided into paired and unpaired datasets, we applied different evaluation metrics for the different datasets. For the datasets with paired images, including LOL and MIT-Adobe FiveK, we employed the PSNR, SSIM, and \( \Delta E \) metrics, where \( \Delta E \) represents the color difference and is formulated such that

\[
\Delta E = \left( \Delta L^2 + \Delta a^2 + \Delta b^2 \right)^{1/2},
\]

where \( L, a, \) and \( b \) represent the Lab color space.

| TABLE 1. Ablation study. Quantitative results under different configurations. “W/O” means without. |
|---------------------------------|-----------|-------|-------|
| MARN w/o IA map                | 21.59     | 0.8665| 10.13 |
| MARN w/o \( L_x \)            | 20.33     | 0.8652| 10.26 |
| MARN w/o \( L_{sim} \)        | 21.66     | 0.8633| 9.78  |
| MARN w/o \( L_{detail} \)     | 21.45     | 0.8678| 10.11 |
| MARN w/o \( L_{color} \)      | 22.10     | 0.8718| 9.84  |
| Residual blocks                | 19.34     | 0.8474| 10.18 |
| MARN                            | 22.33     | 0.8763| 9.77  |
FIGURE 4. Visual comparison of different methods on image 1 from the LOL dataset. Red boxes indicate the obvious differences.

FIGURE 5. Visual comparison of different methods on image 2 from the LOL dataset. Red boxes indicate the obvious differences.

FIGURE 6. Visual comparison of different methods on image 3 from the MIT Adobe FiveK dataset. Red boxes indicate the obvious differences.
As shown in Table 2, our method achieves much better results for PSNR, SSIM, and $\Delta E$. For datasets with unpaired images, we employed the NIQE [42] and CIQI [35] metrics. NIQE measures the visual quality in view of the natural scene statistics and is a widely used non-reference metric, where smaller values indicate better visual quality. CIQI is a non-reference image color quality metric; we employed it to measure the naturalness of the colorfulness of the enhanced images, where higher values indicate better color richness. From Table 3, we can see that our method ranks high for all datasets in terms of the NIQE metric. Only Zero-DCE ranks better for DCIM and LIME. In terms of the CIQI metric, our method ranks first for all five datasets, except MEF, where it scores only 0.007 lower than KinD. In summary, our method outperforms all the other methods on the benchmark datasets. We also test the computationally efficient of different methods. Running time is measured on a PC with an Nvidia GTX 1080Ti GPU and Intel Xeon E5-1603 CPU. Table 4 shows the running time of different methods averaged on 30 images of size $400 \times 600 \times 3$. Only Zero-DCE is faster.
TABLE 4. Running time of different methods. The best result is shown in bold, while the second best is shown in underline.

| Methods     | Time cost(s) | Platform       |
|-------------|--------------|----------------|
| SRIE        | 8.436        | Matlab(cpu)    |
| Retinex-net | 0.665        | Tensorflow(gpu) |
| MBLLEN      | 0.143        | Tensorflow(gpu) |
| KinD        | 0.688        | Tensorflow(gpu) |
| Zero-DCE    | 0.003        | Pytorch(gpu)   |
| Ours        | 0.032        | Pytorch(gpu)   |

than our MARN, however, our method can enhance images with higher dynamic range and is better than Zero-DCE in noise suppression. Our MARN is computationally efficient that can basically meet real-time requirements.

V. CONCLUSION

In this paper, we proposed a novel end-to-end model for low-light image enhancement called MARN. Our method combines the IA map with the low-light image as the input for learning an image-to-illumination mapping, which guides the model to predict an accurate inverse illumination map. To improve the generalization capability and learning ability of the model, we introduced a novel multi-scale attention module. Additionally, we aggregated four loss functions to measure the illuminance, details, and color. Of these functions, the color loss function is non-reference and greatly restores the naturalness of the image color. The experimental results show that the proposed method can not only restore the color information and the detail features of low-light images but also process images with a large dynamic range and improve the brightness and contrast. Our experiments with several state-of-the-art methods on benchmark datasets demonstrate the superiority of our method in terms of visual and quantitative results.

REFERENCES

[1] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. H. Romeny, J. B. Zimmerman, and K. Zuiderveld, “Adaptive histogram equalization and its variations,” Comput. Vis., Graph., image Process., vol. 39, no. 3, pp. 355–368, Sep. 1987, doi: 10.1016/S0734-189X(87)80186-X.

[2] E. D. Pisano, S. Zong, B. M. Hemminger, M. DeLuca, R. E. Johnston, K. Müller, M. P. Braeuning, and S. M. Pizer, “Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms,” J. Digit. Imag., vol. 11, no. 4, pp. 193–200, Nov. 1998, doi: 10.1007/BF03178082.

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

[4] E. H. Land, “The retinex theory of color vision,” Sci. Amer., vol. 237, no. 6, pp. 108–128, Dec. 1977.

[5] Z. Rahman, D. J. Jobson, and G. A. Woodell, “Multi-scale retinex for color image enhancement,” in Proc. 3rd IEEE Int. Conf. Image Process., vol. 3, Sep. 1996, pp. 1003–1006.

[6] D. J. Jobson, Z. Rahman, and G. A. Woodell, “A multiscale retinex for bridging the gap between color images and the human observation of scenes,” IEEE Trans. Image Process., vol. 6, no. 7, pp. 965–976, Jul. 1997, doi: 10.1109/83.597727.

[7] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising,” IEEE Trans. Image Process., vol. 26, no. 7, pp. 3142–3155, Jul. 2017, doi: 10.1109/TIP.2017.2662206.

[8] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, “Toward convolutional blind denoising of real photographs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 1712–1722.

[9] K. Zhang, W. Zuo, and L. Zhang, “FFDDNet: Toward a fast and flexible solution for CNN-based image denoising,” IEEE Trans. Image Process., vol. 27, no. 9, pp. 4608–4622, Sep. 2018, doi: 10.1109/TIP.2018.2839891.

[10] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, Feb. 2016, doi: 10.1109/TPAMI.2015.2439281.

[11] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1646–1654.

[12] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, “Revisiting RCNN: On awakening the classification power of faster RCNN,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 453–468.

[13] J. Redmon and A. Farhadi, “YOLOv3: An incremental improvement,” 2018, arXiv:1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767.

[14] W. Wang, C. Wei, W. Yang, and J. Liu, “GLADNet: Low-light enhancement network with global awareness,” in Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), May 2018, pp. 751–755.

[15] L. Tao, C. Zhu, G. Xiang, Y. Li, H. Jia, and X. Xie, “LLCNN: A convolutional neural network for low-light image enhancement,” in Proc. IEEE Vis. Commun. Image Process. (VCIP), Dec. 2017, pp. 1–4.

[16] C. Wei, W. Wang, W. Yang, and J. Liu, “Deep retinex decomposition for low-light enhancement,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2019, Art. no. 61772043.

[17] Y. Zhang, J. Yang, and X. Guo, “Kindling the darkness: A practical low-light image enhancer,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 1632–1640.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. Int. Conf. Neural Inf. Process. Syst. Cambridge, MA, USA: MIT Press, 2014, pp. 2672–2680.

[19] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, “Zero-reference deep curve estimation for low-light image enhancement,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 1780–1789.

[20] X. Guo, Y. Li, and H. Ling, “LIME: Low-light image enhancement via illumination map estimation,” IEEE Trans. Image Process., vol. 26, no. 2, pp. 982–993, Feb. 2017, doi: 10.1109/TIP.2016.259450.

[21] S. Wang, J. Zheng, H.-M. Hu, and B. Li, “Naturalness preserved enhancement algorithm for non-uniform illumination images,” IEEE Trans. Image Process., vol. 22, no. 9, pp. 3538–3548, Sep. 2013, doi: 10.1109/TIP.2013.2261309.

[22] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, “A weighted variational model for simultaneous reflectance and illumination estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2782–2790.

[23] K. G. Lore, A. Akintayo, and S. Sarkar, “LLNet: A deep autoencoder approach to natural low-light image enhancement,” Pattern Recognit., vol. 61, pp. 650–662, Jan. 2017, doi: 10.1016/j.patcog.2016.06.008.

[24] C. Chen, Q. Chen, J. Xu, and V. Koltun, “Learning to see in the dark,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 3291–3300.

[25] F. Lv, F. Lu, J. Wu, and C. Lim, “MBLLEN: Low-light image/video enhancement using CNNs,” in Proc. Brit. Mach. Vis. Conf. (BMVC), 2018, p. 220.

[26] Y. Jiang, X. Gong, D. Liu, Y. Cheng, and C. Fang, “EnlightenGAN: Deep light enhancement without paired supervision,” IEEE Trans. Image Process., vol. 30, pp. 2340–2349, Jan. 2021, doi: 10.1109/TIP.2021.3051462.

[27] V. Boykovsky, S. Paris, E. Chan, and F. Durand, “Learning photographic global tonal adjustment with a database of input/output image pairs,” in Proc. CVPR, Providence, RI, USA, Jun. 2011, pp. 97–104, doi: 10.1109/CVPR.2011.5995413.

[28] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, “Under-exposed photo enhancement using deep illumination estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 6849–6857.
X. Zhang, X. Wang: MARN for Low-Light Image Enhancement

[29] W. Huang, Y. Zhu, and R. Huang, “Low light image enhancement network with attention mechanism and retinex model,” IEEE Access, vol. 8, pp. 74306–74314, 2020, doi: 10.1109/ACCESS.2020.2988767.

[30] X. Ke, W. Lin, G. Chen, Q. Chen, X. Qi, and J. Ma, “EDLLIE-Net: Enhanced deep convolutional networks for low-light image enhancement,” in Proc. IEEE 5th Int. Conf. Image, Vis. Comput. (ICIVC), 2020, pp. 59–68.

[31] J. Cai, S. Gu, and L. Zhang, “Learning a deep single image contrast enhancer from multi-exposure images,” IEEE Trans. Image Process., vol. 27, no. 4, pp. 2049–2062, Apr. 2018, doi: 10.1109/TIP.2018.2794218.

[32] D. Marnerides, T. Bashford-Rogers, J. Hatchett, and K. Debattista, “ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content,” Comput. Graph. Forum, vol. 37, no. 2, pp. 37–49, May 2018, doi: 10.1111/cgf.13340.

[33] J. Cai, S. Gu, and L. Zhang, “Learning a deep single image contrast enhancer from multi-exposure images,” IEEE Trans. Image Process., vol. 27, no. 4, pp. 2049–2062, Apr. 2018, doi: 10.1109/TIP.2018.2794218.

[34] D. Marnerides, T. Bashford-Rogers, J. Hatchett, and K. Debattista, “ExpandNet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content,” Comput. Graph. Forum, vol. 37, no. 2, pp. 37–49, May 2018, doi: 10.1111/cgf.13340.

[35] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, “Fast and accurate image super-resolution with deep Laplacian pyramid networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 11, pp. 2599–2613, Nov. 2019, doi: 10.1109/TPAMI.2018.2865304.

[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004, doi: 10.1109/TIP.2003.819861.

[37] K. Panetta, C. Gao, and S. Agaian, “No reference color image contrast and quality measures,” IEEE Trans. Consum. Electron., vol. 59, no. 3, pp. 643–651, Aug. 2013, doi: 10.1109/TCE.2013.6626251.

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[39] C. Lee, C. Lee, and C.-S. Kim, “Contrast enhancement based on layered difference representation,” in Proc. 19th IEEE Int. Conf. Image Process., Sep. 2012, pp. 965–968.

[40] M. Jaderberg, K. Simonyan, and A. Zisserman, “Spatial transformer networks,” in Proc. Adv. Neural Inf. Process. Syst., vol. 28, 2015, pp. 2017–2025.

[41] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[42] Y. Guo, Y. Lu, R. W. Liu, M. Yang, and K. T. Chui, “Low-light image enhancement with regularized illumination optimization and deep noise suppression,” IEEE Access, vol. 8, pp. 145297–145315, 2020, doi: 10.1109/ACCESS.2020.3015217.

XIN ZHANG received the B.S. degree from the Beijing Institute of Technology, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Optics and Photonics. His research interests include image restoration, including image denoising and low-light image enhancement.

XIA WANG received the Ph.D. degree in automation from the China University of Mining and Technology, in 1999. She is currently an Associate Professor with the Beijing Institute of Technology, where she is also the Vice Dean of the Institute of Photoelectric Imaging and Information Engineering. Her current research interests include optoelectronic detection, spectrum analysis, and imaging technology.