

SMNet: Synchronous Multi-Scale Low Light Enhancement Network With Local and Global Concern

Shideng Lin, Fan Tang[✉], Weiming Dong[✉], *Member, IEEE*, Xingjia Pan[✉], and Changsheng Xu[✉], *Fellow, IEEE*

Abstract—Limited by objectively poor lighting conditions and hardware devices, low-light images with low visual quality and low visibility are inevitable in the real world. Accurate local details and reasonable global information play their essential and distinct roles in low-light image enhancement: local details contribute to fine textures, while global information is critical for a proper understanding of the global brightness level. In this article, we focus on integrating local and global aspects to achieve high-quality low-light image enhancement by proposing the synchronous multi-scale low-light enhancement network (SMNet). A synchronous multi-scale representation learning structure and a global feature recalibration module are adopted in SMNet. Different from the traditional multi-scale feature learning architecture, SMNet carries out the multi-scale representation learning in a synchronous way: we first calculate the rough contextual representations in a top-down manner and then learn multi-scale representations in a bottom-up way to generate representations with rich local details. To acquire global brightness information, a global feature recalibration module (GFRM) is applied after the synchronous multi-scale representations to perceive and exploit proper global information by global pooling and projection to recalibrate channel weights globally. The synchronous multi-scale representation and GFRM compose the basic local-and-global block. Experimental results on mainstream real-world dataset LOL and synthetic dataset MIT-Adobe FiveK show that the proposed SMNet not only leads the way on objective metrics (0.41/2.31 improvement of PSNR on two datasets) but is also superior in subjective comparisons compared with typical SoTA methods.

Index Terms—Low-light image enhancement, multi-scale feature learning, deep-learning.

Manuscript received 13 December 2021; revised 3 November 2022; accepted 1 March 2023. Date of publication 8 March 2023; date of current version 15 December 2023. This work was supported in part by Beijing Natural Science Foundation under Grant L221013, and National Science Foundation of China under Grants U20B2070, 61832016, 62102162, 62172126, and 62106063. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Mea Wang. (*Corresponding author: Fan Tang.*)

Shideng Lin, Weiming Dong, and Changsheng Xu are with the School of Artificial Intelligence, University of Chinese Academy of Sciences & NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 101408, China (e-mail: linshideng2019@ia.ac.cn; weiming.dong@ia.ac.cn; csxu@nlpr.ia.ac.cn).

Fan Tang is with the Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China (e-mail: tangfan@ict.ac.cn).

Xingjia Pan is with the Youtu Lab, Tencent, Shanghai 200001, China (e-mail: xjia.pan@gmail.com).

The code had already been uploaded to <https://github.com/linshideng/SMNet>.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TMM.2023.3254141>, provided by the authors.

Digital Object Identifier 10.1109/TMM.2023.3254141

I. INTRODUCTION

DUE to the poor objective lighting conditions and the limitation of camera equipment, low-light images in the real world are unavoidable. Low-light images show the characteristics of low overall brightness and image quality with severe degradation, such as noise and blur, which cannot fully express the information in natural shooting scenes, thus limiting the practical application value of the images. Hardware-based approaches, such as using flashlights or increasing exposure time, are limited by usability and performance. Traditional postprocessing low-light image enhancement methods include gamma correction [1], histogram equalization [2], [3], Retinex theory [4], [5], [6], [7], [8], etc. Gamma correction-based methods manually design nonlinear pixel-intensity formulas to change the brightness of images. However, a significantly changed gamma coefficient will lead to a cliff-like drop in brightness. Histogram equalization-based methods extend the original histogram of the image into a new high dynamic range and uniformly transform the histogram to indirectly alter the brightness. However, those methods are usually ineffective with images with a high dynamic range originally. Retinex-based methods decouple low-light images into reflections containing intrinsic contents and illumination components representing brightness information. By enhancing the illumination map, the Retinex-based methods can obtain normal-light images, but the quality of the results is still limited by manual decoupling and enhancement formulas.

Recently, researchers have focused on adopting deep learning frameworks for low-light image enhancement [10], [11]. Lore et al. [12] proposed the first encoder-decoder-based low-light enhancement network by encoding low-light images into hidden space to obtain strong contextual expression ability and decoding them to obtain normal-light images. Afterward, Retinex theory was incorporated into the encoder-decoder pipeline to regulate the pixel-wise generation process with additional domain knowledge [13], [14]. With Retinex decomposition, RetinexNet [9] is by nature a good structure for light condition adjustment and can significantly improve the light condition and recover bright colors. However, the generated visual results are unsatisfactory in terms of both brightness and naturalness (as shown in Fig. 1(b)). To utilize large-scale unpaired data, EnlightenGAN [15] adopted global-local discriminators and additional attention modules across different layers in the encoder-decoder framework to

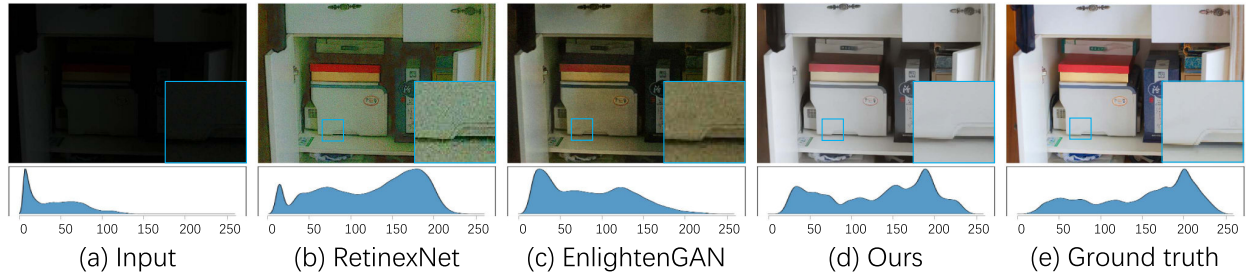


Fig. 1. Visual results of different SoTA methods. Compared with other SoTA methods, our method produces results with better global brightness level and fine local textures without noise on the mainstream real-world dataset LOL [9]. Please zoom in for a better look.

generate natural results. With the global-local discriminators, the overall brightness is more natural than RetinexNet. As shown in Fig. 1(c), the brightness distribution of RetinexNet is more similar to that of the ground truth compared with EnlightenGAN.

However, all of these encoder-decoder-based generators suffer from the missing details in the enhanced results (see blue boxes in Fig. 1(b) and (c)). The frequent down-and-up sampling in the feature flow of encoder-decoder structures leads to the misalignment of local spatial details among input and output images. As a typical pixel-to-pixel generation task, low-light image enhancement requires accurate local details and reasonable global brightness information to maintain the content consistency of the input and the output and improve the brightness level of the output.

Considering the importance of accurate local details and global brightness distribution in low-light conditions, we propose a synchronous multi-scale representation network (SMNet) for learning effective feature representations. Different from traditional multi-scale learning approaches [10], [16] where the multi-scale features are learned in a parallel way, the proposed synchronous multi-scale feature learning framework arranges the feature flow of different scales in a top-down to bottom-up way. The low-resolution bottom branch first extracts high-level information in the top-down flow. The top high-resolution branch of the multi-scale network without redundant up-and-down sampling is guided to generate accurate local details by residual links across different branches as compensation. The importance of emphasizing the computation order lies in the fact that the distribution of degraded information, such as noise, is closely related to contextual information. Then, features from different scales are concatenated as input to a global feature recalibration module (GFRM) that recalibrates the multi-scale features from the channel perspective. As shown in Fig. 1(d), the proposed SMNet could generate high-quality results with natural visual effects and local details.

Our main contributions can be summarized as follows:

- We propose the synchronous multi-scale network (SMNet) by stacking several L&G blocks that contain a novel top-down to bottom-up feature flow in the synchronous multi-scale structure to obtain the high-quality outputs of low-light image enhancement.
- We propose a global feature recalibration module (GFRM) to perceive global information and recalibrate weights to keep the proper global brightness information.

- Extensive experimental results show that the proposed SMNet achieves state-of-the-art performance in mainstream real-world dataset LOL and synthetic dataset MIT-Adobe FiveK.

II. RELATED WORKS

A. Low-Light Image Enhancement

LL-Net [12] was the first application to enhance low-light images using a deep learning-based architecture. Considering noise in low-light images, the LL-Net stacked sparse denoising automatic encoder realizes normal-light images with noise suppression. Combined with the classical Retinex theory, which decomposes images into reflection components and illumination components, many learning-based methods [9], [14], [17], [18] have been proposed, among which DeepUPE [14] is a typical representative. DeepUPE only learns an illumination-to-image mapping function by a classic encoder-decoder structure. Compared with other Retinex-based methods considering the computation of the reflectance component, DeepUPE has less computation in processing low-light images. Focusing on the enhancement of extremely low-light images, EEMEFN [19] brightened low-light images while trying to maintain the sharp edges of the inputs. Adopting raw images as input, which is different from others using JPEG, Chen et al. [20] proposed a framework that can achieve a raw-to-JPEG enhancement. Benefiting from the ability of unsupervised learning of Generative Adversarial Networks (GANs) [21], Jiang et al. [15] introduced a GAN-based method called EnlightenGAN to train with unpaired datasets. Wang et al. [22] proposed DLN, which incorporates knowledge of back projection [23] to learn a residual feature map to lighten low-light images. TBEFN [24] is a two-branch exposure-fusion Retinex-based network for images with different illumination conditions. Considering the importance of perceptual quality, DRBN [11] introduced a two-stage framework that learns a band representation to obtain basic normal-light images in the first stage. In the second stage, DRBN recomposed the band representation to improve the perceptual quality of basic normal-light images via adversarial learning. By learning an optimal curve-to-image function, ZERO-DCE [25] obtained enhanced results training with unpaired datasets. Zamir et al. [10] proposed a multi-scale framework called MIRNet, which realized different low-level tasks, including low-light image enhancement by maintaining spatially-precise high-resolution representations.

Retinex theory-based methods use reflection components and illumination components with limited nonlinear expression ability. In contrast, others adopt the structure of an encoder-decoder that may cause artifacts due to frequent up-and-down sampling. Our method emphasizes the reasonable top-down and bottom-up mobility of complementary feature information in the proposed synchronous multi-scale structure, which can ensure accurate local spatial details. In addition, we also pay attention to the global brightness information using GFRM and try to restore the information of the real shot scene.

B. Multi-Scale Representation Learning

For deep learning-based methods, the structure design is the key point that affects the performance of a certain model [26]. Plain network [26], [27], [28], [29], encoder-decoder [12], [30], [31], multi-stages [11], [32] and multi-scales [33], [34], [35], [36], [37] are representative architectures in deep learning-based methods. Multi-scale network structures have achieved good results in many computer vision tasks [38], [39], [40], including low-level tasks, such as image deblurring [34], [41], deraining [42], dehazing [35], and low-light image enhancement [10]. EDPN [41] is a multi-scale method for blurry image restoration from multiple degradations, and fully exploiting the self- and cross-scale similarities in the degraded image by two pyramid-based modules. Tao et al. [34] proposed a multi-scale model SRN-DeblurNet that recurrently utilizes features in high dimensions. Still, the SRN-DeblurNet pays too much attention to high-dimensional features, leading to poor performance in the task of image enhancement or generation, where spatial information is also essential. LPNet [42] is a lightweight structure based on a multi-scale Gaussian-Laplacian pyramid network with clear interpretability, but it is difficult to converge because of computing loss at every single scale. Liu et al. [35] presented a grid net named GridDehazeNet, which features flow and connects freely at different scales. MIRNet [10] is a multi-scale framework that realizes different low-level tasks, including low-light image enhancement. However, this complicated and nonfocal point network structure design cannot guarantee that useful information to be spread in a reasonable way.

Different from other multi-scale structure methods, considering the pixel-to-pixel modification task of low-light image enhancement, our method first calculates the features in the low-resolution high-dimension branch to extract high-level contextual information and then moves on to the calculation of spatial detail features in a nonparallel way. The information from the low-resolution high-dimension branch serves as a guide in a top-down way for the high-resolution low-dimension branch, which is related to the generation of spatial details. Reasonable design of multi-scale structure helps us to obtain a high-quality output with fine texture.

C. Channel-Wise Feature Recalibration

The weight distribution of feature information has an obvious influence on network performance. To investigate the relationship between performance and distribution of weights, many works [43], [44], [45], [46], [47], [48] have been proposed. SENet [44] was proposed to focus on the distinction between

channels by first using global average pooling to obtain the channel-squeeze-weight and then recalibrating these weights to enhance the feature hierarchy. To thoroughly investigate the channel dependency, Zhang et al. [45] proposed RCAN with a residual channel attention block and short residual connection. This structure can strengthen the representational ability of CNNs and discriminate the importance of different channels. Based on SENet [44], CBAM [46] considered spatial information in addition to channel attention and discussed the need for additional global max pooling use. The channel module and spatial module are connected in series. SRM [47] performed style pooling and style integration on images in series, essentially aiming to recalibrate the weights of crucial style information of images in the task of neural style transfer.

Our proposed SMNet emphasizes the importance of different hierarchical features in channels, especially for the multi-scale features that already contain rich local information. By combining the global average pooling, global max pooling, and a fully connected layer in parallel, we can recalibrate the weights adaptively in an effective and efficient way.

D. Denoising

In low-light image enhancement, the input carries the noise degradation information, which reduces the output quality. Traditional classical denoising methods, such as BM3D [49] and NLM [50] have good denoising efficiency, but the methods based on deep learning have better performance, such as DnCNN [51], RED [52] and SOD-CED [53]. Based on these deep denoising models, researchers also proposed a self-supervised method Neighbor2Neighbor [54], that can complete the training without clean images. Combining denoising with Neural Architecture Search (NAS) [55], NAS-DIP [56] can autonomously search a reasonably deep architecture to produce a clean image from noisy images. In the field of low-light image enhancement, researchers prefer to modularize the denoising ability, like the noise encoder-decoder structure in LLNet [12]. In this paper, we do not pay special attention to reduce the influence of noise information but focus on enhancing the low-light inputs in an end-to-end fashion.

Finally, we summarised the low-light image enhancement methods and denoising methods in Table I.

III. METHODOLOGY

A. Overall Structure

Recovering high-quality image content from low-light input is challenging since the image captured under low-light conditions is usually dim and noisy. Desired high-quality results of low-light image enhancement should not only have an appropriate brightness level but also have a clear fine texture while eliminating inherent degradation such as noise and blur. To this end, we propose a synchronous multi-scale low light enhancement network (SMNet) to take into account local details. A global feature recalibration module (GFRM) is further proposed to extract global information.

The overall structure of SMNet is shown in Fig. 2. Following Zhang et al. [18], we combine the low-light RGB image $\mathbf{I}_{in}$

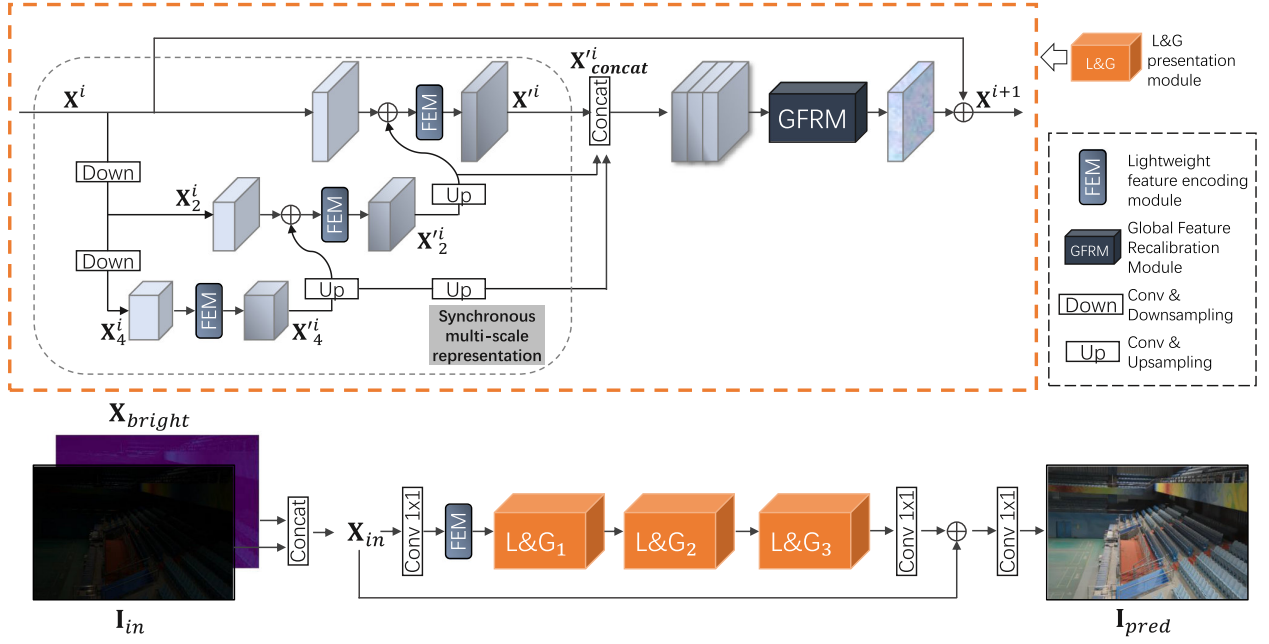


Fig. 2. The architecture of the proposed synchronous multi-scale low light enhancement network (SMNet) mainly consists of local-global (L&G) blocks for low-light image enhancement. With the help of the lightweight feature encoding module (FEM), the multi-scale network provides accurate local details. Moreover, the proposed global feature recalibration module (GFRM) can perceive complementary global information and adaptively recalibrate the useful features.

TABLE I
SUMMARY OF REPRESENTATIVE LOW-LIGHT IMAGE ENHANCEMENT METHODS
AND DENOSING METHODS

Method	Kind	Year	Published
Gamma Correction	traditional	NA	NA
AHE	traditional	1987	CVGIP
CLAHE	traditional	1998	JDI
SSR	traditional	1997	TIP
MSR	traditional	1996	ICIP
MSRCR	traditional	1997	TIP
LIME	traditional	2016	TIP
JieP	traditional	2017	CVPR
DUAL	traditional	2019	CGF
LLNet	deep learning	2017	PR
RetinexNet	deep learning	2018	BMVC
Chen et al.	deep learning	2018	CVPR
KIND	deep learning	2019	MM
DeepUPE	deep learning	2019	CVPR
EEMEFN	deep learning	2020	AAAI
DLN	deep learning	2020	TIP
DRBN	deep learning	2020	CVPR
EnlightenGAN	deep learning	2021	TIP
ZERO-DCE	deep learning	2020	CVPR
MIRNet	deep learning	2020	ECCV
TBEFN	deep learning	2021	TMM
NLM	traditional	2005	CVPR
BM3D	traditional	2017	TIP
RED	deep learning	2016	NIPS
DnCNN	deep learning	2017	TIP
SOD-CED	deep learning	2019	IET CV
NAS-DIP	deep learning	2020	ECCV
Hou et al.	deep learning	2021	CVPR

with the maximum value of RGB channels $\mathbf{X}_{bright}$ serving as input $\mathbf{X}_{in} \in \mathbb{R}^{H \times W \times 4}$ to make the best use of the luminance information. One 1×1 convolution layer is involved to increase the dimension of $\mathbf{X}_{in}$.

Lightweight feature encoding: Before introducing the proposed synchronous multi-scale feature learning mechanism, we utilize a lightweight feature encoding module (FEM) to fuse $\mathbf{X}_{in}$ following [57]. The FEM is light and effective by only using two 3×3 convolution layers followed by the PReLU layer [58] and two 1×1 convolution layers. Different from inverted block [59], the proposed FEM first decreases the dimension and then increases to obtain more expressive extracted information. We use 3×3 convolutions at both ends and 1×1 convolutions at the bottleneck. To avoid loss of spatial local information, there is no up- or down-sampling in the FEM. Furthermore, we establish a shortcut between high dimensions to better optimize network training.

After FEM, we propose the local-global (L&G) block, which consists of a synchronous multi-scale feature learning approach (see Section III-B) and a global feature recalibration module (see Section III-C). By stacking L&G blocks several times, we set up SMNet to consider both local details for fine texture and global information for the global brightness level. Loss functions for training are listed in Section III-D.

B. Synchronous Multi-Scale Feature Learning

Given the input of the i -th L&G module $\mathbf{X}^i$, we conducted a downsampling operation to obtain small-scale feature maps $\mathbf{X}_d^i$, $d = \{2, 4\}$. When $d = 2$, the channel number of $\mathbf{X}_2^i$ becomes double, and the resolution becomes half compared with $\mathbf{X}^i$. When $d = 4$, the modification coefficients become 4 and $1/4$. As shown in Fig. 2 (top), three scale-different branches are set up to enhance the local details for feature learning. Features extracted in high dimensions contain rich contextual information, which is helpful for low-dimensional branches to

error is lower than δ , huber loss is L_2 loss; otherwise, it is L_1 loss. The expression of huber loss is defined as:

$$\begin{aligned} H_{Diff} &= \frac{1}{2}(I_{pred} - I_{gt})^2 \circ M_{|I_{gt}-I_{pred}|<\delta} \\ &\quad + \delta \left(|I_{gt} - I_{pred}| - \frac{1}{2}\delta \right) \circ M_{|I_{gt}-I_{pred}|\geq\delta}, \\ \mathcal{L}_h &= \sum_{i=1}^H \sum_{j=1}^W H_{Diff} \end{aligned} \quad (5)$$

where $\circ$ is the Kronecher product, M_* is the binary mask, and δ is a coefficient that controls the sensitivity to outlier values. We use $\delta = 1$ in our experiment empirically.

4) *TV Loss*: For noise, the undesirable degradation that occurs after the brightness enhancement process should be mitigated in the final results. Based on [61], we know that noisy images are usually accompanied by high variance. It is necessary to reduce the total changes between adjacent pixels to obtain the clean outputs. The TV loss is defined as follow:

$$\begin{aligned} \mathcal{L}_{tv}(I_{pred}) &= \sum_{i=1}^H \sum_{j=1}^W \\ &\quad \sqrt{(I_{pred}(i, j) - I_{pred}(i+1, j))(I_{pred}(i, j) - I_{pred}(i, j+1))}, \end{aligned} \quad (6)$$

where $I_{pred}(i, j)$ denotes the specific pixel value in I_{pred} , and the terms i and j indicate the index of all pixels.

5) *Color Loss*: Inspired by [14], we applied the color loss as the chroma estimator. By calculating the cosine distance between the ground truth and predictions, color loss can reduce the color distortion problem that usually occurs in the enhancement process. It can be formulated as:

$$\mathcal{L}_c = \sum_{i=1}^H \sum_{j=1}^W \angle(I_{pred}(i, j), I_{gt}(i, j)), \quad (7)$$

where $\angle(\cdot, \cdot)$ calculates the angle between two colors regarding the RGB color as a 3D vector.

6) *Total Loss* As a result, the total loss of our model is formulated as:

$$\mathcal{L}_{total} = \omega_s \mathcal{L}_s^i + \omega_p \mathcal{L}_p^i + \omega_h \mathcal{L}_h^i + \omega_c \mathcal{L}_c^i + \omega_{tv} \mathcal{L}_{tv}^i. \quad (8)$$

We practically set ω_s , ω_p , ω_h and ω_c to be 1.0, and $\omega_{tv} = 0.001$. We also carried out an ablation experiment for losses in Section IV-C.

IV. EXPERIMENTS

A. Implementation Details

We employed PyTorch [62] to accomplish our experiments on an Nvidia Titan XP GPU. The initial learning rate was set at 5×10^{-4} , which decreased by 1.2 times every 50 epochs. To enrich the diversity, augmentation operations such as random flip and rotation were applied to the mainstream real-world training dataset LOL [63] and synthetic dataset MIT-Adobe FiveK [64]. During the training phase, we use Adam [65] as the optimizer to optimize all parameters for a total of 600 epochs.

TABLE II
SPEED OF INFERENCE (FRAMES PER SECOND, FPS) OF DIFFERENT METHODS
WHEN THE RESOLUTION OF THE INPUT IMAGE IS 600×400

Methods	FPS	Parameters	FLOPs
RetinexNet	1.90	0.555M	587.470G
EnlightenGAN	21.71	8.636M	273.240G
MIRNet	1.99	31.787M	2988.84G
TBEFN	5.50	0.490M	108.532G
ZERO-DCE	61.23	0.079M	84.990G
Ours	32.40	2.989M	1161.91G

We also show the model size and floating-point operations per second (FLOPs) of different methods.

B. Evaluations

To demonstrate the superiority of our method, we conduct evaluations on datasets of LOL and MIT-Adobe FiveK respectively. For the sake of fairness, all selected methods are trained in the corresponding training set. For the LOL dataset, LIME [66], RetinexNet [9], KinD [17], EnlightenGAN [15], TBEFN [24], DRBN [11], and MIRNet [10] are compared. For the MIT-Adobe FiveK dataset, DUAL [67], JieP [68], DeepUPE [14], ZERO-DCE [25] and MIRNet [10] are compared.

The training configuration for the LOL dataset is the same as that for the DLN [22]. For MIT-Adobe FiveK, we choose the version of Expert C following [14], and then randomly select 4,500 images as the training set. The remaining 500 images are used for validation and testing. Table II shows the comparisons of testing speed, model size, and FLOPs of different methods.

1) *Qualitative Evaluation*: Figs. 4 and 5 show the qualitative comparisons with existing state-of-the-art methods on mainstream real-world dataset LOL and synthetic dataset MIT-Adobe FiveK. Overall, our method can produce perceptually faithful outputs, which not only restore the real scene information well with appropriate brightness level but are also perceptually favorable. At the same time, for local details, our method can generate results with finer textures, which have clearer edges and less degradation, such as noise or color distortion, than other methods.

As shown in Fig. 4, LIME [66] cannot deal with extremely low-light image images regardless of the brightness level or details, making the output darker and full of degradation than other methods. RetinexNet [9] and KinD [17] combine deep learning with Retinex theory and achieves results with a good brightness level, but blur occurs in detail texture. EnlightenGAN [15] generates results with serious color distortion and has difficulty dealing with noise. The results of TBEFN [24] are blurred in detail and have low saturation. DRBN [11] produces fine texture, but it also has shortcomings in brightness level. MIRNet [10] achieves the closest quality to our method in both brightness level and details, but it has difficulty controlling artifacts.

As shown in Fig. 5, the two classical Retinex-based methods, DUAL [67] and JieP [68] produce similar darker and low contrast outputs. DeepUPE [14] can generate results with good brightness contrast, but color distortion may appear (e.g., the girl's hair). ZERO-DCE [25] produces results with low contrast and color distortion with undesirable texture. MIRNet [10]

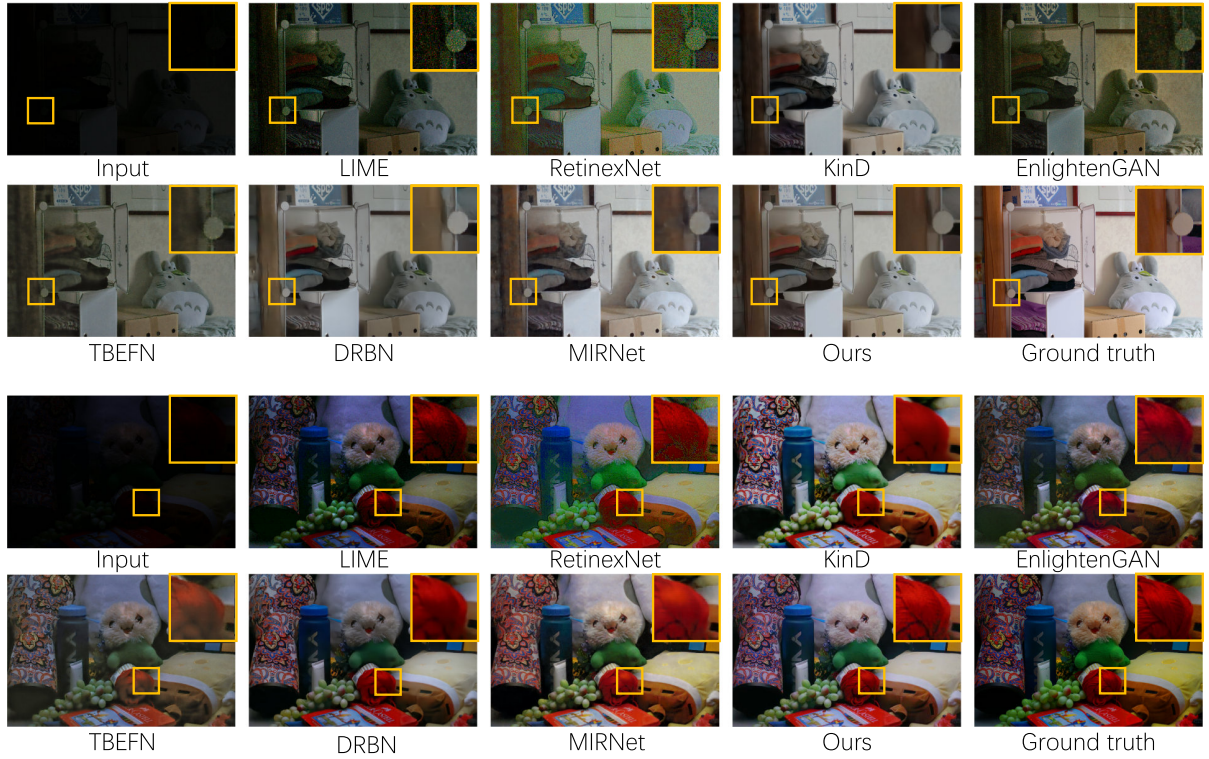


Fig. 4. Qualitative comparisons with state-of-the-art methods on mainstream real-world dataset LOL [9]. Our method produces perceptually faithful output with better texture. The larger yellow box on the top right shows the amplified details of the lower left yellow box. Please zoom in for a better view.



Fig. 5. Qualitative comparison with state-of-the-art methods on the synthetic MIT-Adobe FiveK dataset [64]. Our method produces better texture and perceptually faithful output. Please zoom in for a better view.

achieves the closest quality to our method, but the results occur at warmer color temperatures.

We can see that on those two datasets, even in extremely low-quality conditions, our method can produce visually pleasing results with appropriate global brightness levels and fine local textures. More specifically, our method can effectively avoid noise and color distortion problems while maintaining the consistency and authenticity of the output results with real scenes.

2) *Quantitative Evaluation*: For quantitative comparison, we select the peak signal-to-noise ratio (PSNR) and structural

similarity (SSIM) [69] as objective metrics, which are the mainstream metrics in the field of image enhancement. We also select additional metrics such as information content weighted structural similarity (IW-SSIM) [70], visual information fidelity (VIF) [71], feature similarity index measure (FSIM) [72], and lightness order error (LOE) [4] for a comprehensive comparison. Except for LOE, the other metrics measure the quality with reference ground truth. A higher PSNR score indicates the higher objective quality of an image. A higher SSIM or IW-SSIM score indicates better structural similarity between the ground truth and

TABLE III
QUANTITATIVE COMPARISON ON THE LOL DATASET [9]

Metrics	LIME	RetinexNet	EnlightenGAN	KIND	TBEFN	DRBN	MIRNet	Ours
PSNR $\uparrow$	15.24	16.77	18.23	20.87	17.35	20.13	24.14	24.55
SSIM $\uparrow$	0.470	0.560	0.617	0.802	0.777	<u>0.830</u>	<u>0.830</u>	0.864
IW-SSIM $\uparrow$	0.8768	0.7992	0.8999	0.9065	0.9234	0.9428	<u>0.9521</u>	0.9623
VIF $\uparrow$	0.3890	0.2600	0.3765	0.3960	0.4169	<u>0.4691</u>	0.4331	0.5351
FSIM $\uparrow$	0.9206	0.8592	0.9226	0.9253	0.9438	0.9585	<u>0.9605</u>	0.9690
LOE $\downarrow$	405.74	462.04	395.81	399.81	325.68	480.29	<u>197.35</u>	182.07

The best and second best results are highlighted in bold and underline.

TABLE IV
QUANTITATIVE COMPARISON ON THE MIT-ADOBE FIVEK DATASET [64]

Metrics	DUAL	JieP	DeepUPE	ZERO-DCE	MIRNet	Ours
PSNR $\uparrow$	16.62	17.50	<u>19.92</u>	15.99	19.66	22.23
SSIM $\uparrow$	0.743	0.764	<u>0.803</u>	0.720	0.799	0.849
IW-SSIM $\uparrow$	0.8398	0.8906	<u>0.9424</u>	0.8050	0.9371	0.9561
VIF $\uparrow$	0.3883	0.4976	<u>0.5364</u>	0.4751	0.4144	0.5600
FSIM $\uparrow$	0.8905	0.9158	<u>0.9346</u>	0.8829	0.9293	0.9411
LOE $\downarrow$	374.45	276.23	200.92	376.00	<u>196.50</u>	126.21

The best and second best results are highlighted in bold and underline.

TABLE V
ABLATION STUDY OF L&G PRESENTATION MODULE

Architectures	PSNR $\uparrow$	SSIM $\uparrow$
One L&G block	21.71	0.847
Two L&G blocks	22.84	0.856
SMNet with three L&G blocks	24.55	0.864
W/o multi-scale structure	22.71	0.850
W/o 4 $\times$	22.51	0.844
W/o 2 $\times$ & 4 $\times$	21.73	0.841
Parallel branches	22.58	0.852

The third row is our full model with best performance. The 4 and 2 denote downsample scale branch in multi-scale network of L&G presentation modules. W/o multi-scale structure means removing top-down and bottom-up mechanism.

prediction. By measuring the signal fidelity information between the reference and prediction, a higher VIF score means better fidelity. FSIM is considered the best overall performance metric [73]. The higher the FSIM score is, the more realistic the underlying features in the prediction that conform to the human visual system are. Unlike the other metrics, LOE is a non-reference method. A lower LOE value represents a better naturalness of the output. As shown in Tables III and IV, our method achieved leading performance under all objective metrics on LOL and MIT-Adobe FiveK compared with the state-of-the-art methods.

C. Ablation Study

1) *Effectiveness of the L&G block*: The L&G block is the basic module of the proposed SMNet. We perform an ablation study on the L&G presentation module by following three kinds of attempts: 1) stacking different numbers of L&Gs; 2) removing multi-scale structures (top-down and bottom-up mechanism) in L&Gs; 3) just abandoning the 4 $\times$ and 2 $\times$ branches in the multi-scale representation structure; 4) replacing 4 $\times$ and 2 $\times$ branches by parallel branches.

The quantitative results are shown in Table V. The proposed SMNet with three L&G blocks obtains the best results under both metrics. With the decrease of L&G presentation modules, the performance of the model decreases significantly. One L&G block only achieved PSNR = 21.71 dB and SSIM = 0.847. Two L&G blocks achieved PSNR = 22.84 dB and SSIM = 0.856. The reason is that when removing multi-scale structure in L&G, which means features in the three branches are calculated without interdependence, the asynchronous structure loses reasonable feature flow, and the features extracted from different scales cannot communicate with each other. Similarly, removing 4 $\times$ and/or 2 $\times$ branches will make the network context unaware and lead to a drop in performance. The same result happened when we replacing 4 $\times$ and 2 $\times$ branches by parallel branches.

To further investigate the rationality of the proposed synchronous multi-scale representation structure, we show qualitative comparison results in Fig. 6. In the ground truth image, there are tiny granite-like textures in the top-right region of the wall (in the yellow box). However, the textures in TBEFN and MIRNet are intermingled and lose details, while the results of EnlightenGAN are seriously noisy and cannot even distinguish between real textures or noise. As shown in Fig. 6(e), although degradation of information such as noise is not involved, the result is excessively smooth and loses the granite-like details without the top-down and bottom-up feature flow. On the other hand, SMNet can preserve tiny granite-like textures with sharp edges and suppressed noise. The above results show that the proposed synchronous multi-scale representation structure is objective and visually effective in low-light image enhancement to produce high-image-quality outputs.

2) *Effectiveness of FEM*: A lightweight feature encoding module is used in the multi-scale feature learning pipeline. We stack four regular 3 $\times$ 3 convolutions and pReLU for ablation of the FEM to explore its influence on the performance of the model. None 1 $\times$ 1 convolutions is involved like FEM. In addition,

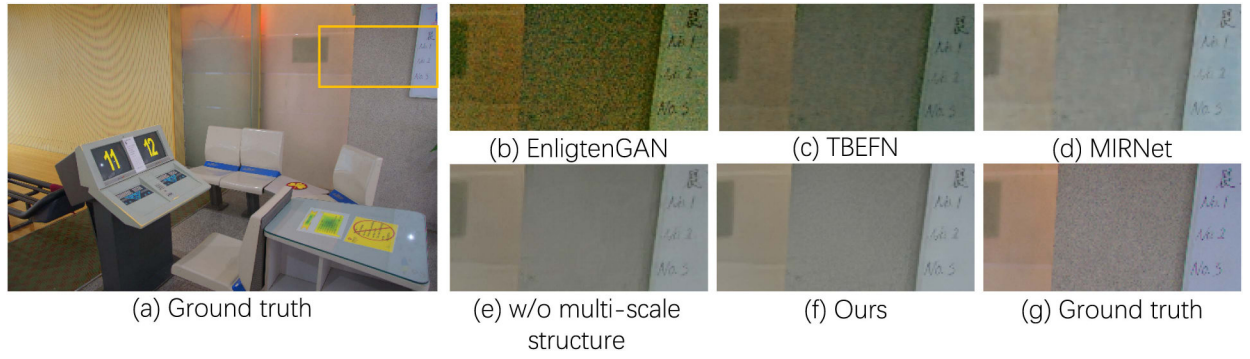


Fig. 6. Visual results to prove our L&G presentation module and multi-scale structure (top-down and bottom-up mechanism) are objective and visually effective to produce high image-quality outputs. Yellow box is the enlarged area.

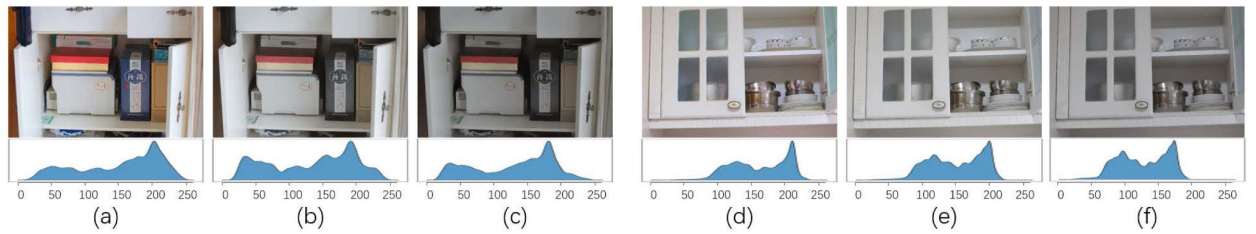


Fig. 7. Ablation study of GFRM. (a) and (d) are ground truth. (b) and (e) are results of SMNet. (c) and (f) are the results of SMNet without GFRM, which are darker, and the distribution of the histogram is more to the left.

the number of feature channels is not changed, and the size of the feature map remains the same. Without FEM, the PSNR of the result drops to 22.98 dB and the SSIM drops to 0.844. By the 1×1 convolutions, the FEM encodes abundant information from the channel changes and improves the nonlinear ability of the current branch. The FEM is also a lightweight module that can be embedded in various branches of multi-scale networks for reuse. Without redundant up-and-down sampling, FEM, to some extent, will not cause the loss of spatial local information.

3) *Effectiveness of GFRM*: To prove the effectiveness of our GFRM in exploiting global information, we perform the ablation of GFRM by replacing or changing the structure of GFRM. We ablated GFRM with stacked convolutions with pReLU. Fig. 7 shows the qualitative results of SMNet with and without GFRM. The results of SMNet without GFRM are darker, and the distribution of the histogram is more to the left. Quantitative results are shown in Table VI. Overall, with the help of GFRM, the results are brighter and closer to the ground truth. The histograms of SMNet with GFRM are also almost the same as the ground truth. Two global pooling operations are adopted in GFRM. To investigate the effectiveness of using two branches simultaneously, we abandon the GAM and GMP, respectively, in the ablation study. The statistics are shown in Table VI. We can see that the performance will be degraded when either GMP or GAP is removed. GMP can produce the most significant features. However, when considering the global relationship of all spatial pixels, GMP leads to the loss of spatial local information. GAP considers the overall relationship of spatial local pixels but may contain information irrelevant to specific tasks. In fact, GMP and GAP can jointly produce complementary information to help improve the performance of the model. For further prove the effectiveness

TABLE VI
ABLATION STUDY OF GFRM

Architectures	PSNR $\uparrow$	SSIM $\uparrow$
W/o GFRM	19.87	0.833
GFRM w/o GAP	22.16	0.845
GFRM w/o GMP	23.26	0.845
GFRM replaced by SENet	21.48	0.844
GFRM replaced by RCAN	23.16	0.855
GFRM replaced by ECANet	23.31	0.849
GFRM replaced by CBAM	22.71	0.842
GFRM replaced by Hou <i>et al.</i>	19.83	0.839
GFRM replaced by SRM	23.03	0.856
SMNet	24.55	0.864

Our full model achieves the best performance.

TABLE VII
ABLATION STUDY OF LOSS CONFIGURATION

Loss Function	PSNR $\uparrow$	SSIM $\uparrow$
W/o $\mathcal{L}_s$	22.53	0.839
W/o $\mathcal{L}_p$	22.65	0.843
W/o $\mathcal{L}_h$	22.52	0.854
W/o $\mathcal{L}_c$	22.31	0.854
W/o $\mathcal{L}_{tv}$	23.06	0.859
SMNet	24.55	0.864

Our full model achieves the best performance.

of GFRM, Table VII shows the ablation experiment results of replacing SOTA attention mechanisms of GFRM. SENet [44] is a classical channel attention mechanism. Through GAP, it obtains the characteristics of different channel layers. RCAN [45] adaptively rescales channel-wise features by considering inter-dependencies among channels. ECANet [74] proposes a local

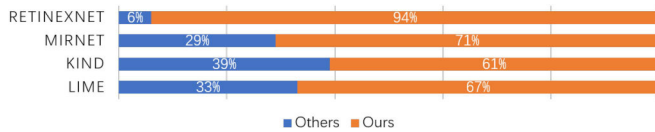


Fig. 8. User study. Compared with other state-of-the-art methods, our method always has a higher selection rate.

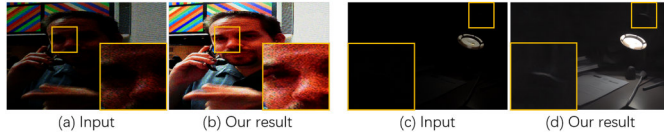


Fig. 9. Bad cases. Our method fails to enhance the low-light images with (a) exaggerated noise or (c) ultra-dark areas.

cross-channel interaction strategy without dimensionality reduction and a method to adaptively select kernel size of 1D convolution, determining coverage of local cross-channel interaction. In addition to channel attention, CBAM [46] and coordinate also considered spatial location information. First, they obtain the channel information through global pooling, and obtain the 2d spatial attention diagram through convolution, to improve the attention to the spatial location information and achieve the improvement of efficiency. In addition, Hou et al. [75] also divides the channel attention into two parallel one-dimensional feature coding processes to effectively integrate the spatial coordinate information into the generated attention map. SRM [47] implements style pooling in spatial dimensions and combines feature maps with style pooling results. Our GFRM perceives and exploit proper global information by global pooling and projection to recalibrate channel weights globally.

Overall, GFRM perceives and exploits global information by global pooling and projection to recalibrate channel weight to obtain proper global brightness level information.

4) *Configuration of Losses*: We perform an ablation study to investigate the impact of losses in our model. Our full model obtains the best performance, as shown in Table VII. When SSIM loss $\mathcal{L}_s$ is deprecated, the SSIM score drops to the lowest. The reason is that $\mathcal{L}_s$ controls the structural similarity of images, which plays a key role in ensuring higher-quality output. Without perceptual loss, $\mathcal{L}_p$, scores of metrics decrease due to a lack of high-level contextual information. The Huber loss $\mathcal{L}_h$ can help to smooth the image details. The absence of $\mathcal{L}_h$ and $\mathcal{L}_c$ has an adverse effect on the performance of the proposed method. TV loss $\mathcal{L}_{tv}$ is mainly used for noise reduction. The noise in the low-light images will be amplified after enhancement, so eliminating properly used $\mathcal{L}_{tv}$ will lead to performance degradation.

D. User Study

Subjective perception is also a critical metric that should be taken into account in the task of low-light image enhancement. Therefore, we conducted a user study of 39 randomly selected participants involving four state-of-the-art methods: LIME [66], RetinexNet [9], KinD [17] and MIRNet [10]. We present the results of SMNet and another randomly selected method on the

same page (two images on one page) to the participants and ask them to choose which one has a higher overall quality and detail quality. As shown in Fig. 8, our method always has a higher selection rate than any other method, which proves that our results are better in subjective perception.

E. Limitations

Although our method can achieve good results in most cases, it still has limited ability in extreme cases, such as dealing with images with exaggerated noise or ultra-dark areas. Bad cases are shown in Fig. 9. For the exaggerated noise case, the enhancement results retain most of the global brightness. However, the detailed textures cannot be recovered well. When dealing with an input image with ultra-dark areas, our method cannot enhance those areas well, because the information on these areas is completely lost.

V. CONCLUSION

We propose a synchronous multi-scale low-light enhancement network (SMNet), which can reasonably increase the brightness of the low-light images and realize the accurate mapping of the content on input and output to obtain fine texture. Special attentions are paid to the global brightness distribution and local details captured in low-light conditions. A synchronous multi-scale network is proposed with carefully designed feature flow to enhance the interaction between different resolutions. Furthermore, a global feature recalibration module is applied after the synchronous multi-scale representations to perceive and exploit proper global information. Experimental results on mainstream datasets show that our method is superior to other existing methods. In future work, we plan to extend our method to low-light video enhancement.

REFERENCES

- [1] S.-C. Huang, F.-C. Cheng, and Y.-S. Chiu, "Efficient contrast enhancement using adaptive gamma correction with weighting distribution," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1032–1041, Mar. 2013.
- [2] S. M. Pizer et al., "Adaptive histogram equalization and its variations," *Comput. Vis., Graph., Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.
- [3] E. D. Pisano et al., "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *J. Digit. Imag.*, vol. 11, no. 4, 1998, Art. no. 193.
- [4] 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.
- [5] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new low-light image enhancement algorithm using camera response model," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 3015–3022.
- [6] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Trans. Image Process.*, vol. 6, no. 3, pp. 451–462, Mar. 1997.
- [7] D. J. Jobson, Z.-u. 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.
- [8] C.-H. Lee, J.-L. Shih, C.-C. Lien, and C.-C. Han, "Adaptive multiscale retinex for image contrast enhancement," in *Proc. Int. Conf. Signal-Image Technol. Internet-Based Syst.*, 2013, pp. 43–50.
- [9] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. 29th Brit. Mach. Vis. Conf.*, 2018, pp. 1–12.
- [10] S. W. Zamir et al., "Learning enriched features for real image restoration and enhancement," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 492–511.

- [11] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3063–3072.
- [12] 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, 2017.
- [13] J. Liang et al., "Deep bilateral retinex for low-light image enhancement," 2020, *arXiv:2007.02018*.
- [14] R. Wang et al., "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6849–6857.
- [15] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.
- [16] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1448–1457.
- [17] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1632–1640.
- [18] Y. Zhang, X. Di, B. Zhang, and C. Wang, "Self-supervised image enhancement network: Training with low light images only," 2020, *arXiv:2002.11300*.
- [19] M. Zhu, P. Pan, W. Chen, and Y. Yang, "Eemefn: Low-light image enhancement via edge-enhanced multi-exposure fusion network," in *Proc. 34th AAAI Conf. Artif. Intell.*, 2020, pp. 13106–13113.
- [20] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3291–3300.
- [21] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [22] L.-W. Wang, Z.-S. Liu, W.-C. Siu, and D. P. Lun, "Lightening network for low-light image enhancement," *IEEE Trans. Image Process.*, vol. 29, pp. 7984–7996, 2020.
- [23] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1664–1673.
- [24] K. Lu and L. Zhang, "TBEFN: A two-branch exposure-fusion network for low-light image enhancement," *IEEE Trans. Multimedia*, vol. 23, pp. 4093–4105, 2021.
- [25] C. Guo et al., "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1780–1789.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–14.
- [28] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [29] K. Kumar and D. D. Shrimankar, "Deep event learning boost-up approach: DELTA," *Multimedia Tools Appl.*, vol. 77, no. 20, pp. 26635–26655, 2018.
- [30] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
- [31] X. Xu, S. Wang, Z. Wang, X. Zhang, and R. Hu, "Exploring image enhancement for salient object detection in low light images," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1s, pp. 1–19, Mar. 2021.
- [32] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5978–5986.
- [33] S. Nah, T. Hyun Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3883–3891.
- [34] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8174–8182.
- [35] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7314–7323.
- [36] A. Zhang, X. Jiang, B. Zhang, and X. Cao, "Multi-scale supervised attentive encoder-decoder network for crowd counting," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 1s, pp. 1–20, Mar. 2020.
- [37] S. Lim and W. Kim, "DSLRL: Deep stacked Laplacian restorer for low-light image enhancement," *IEEE Trans. Multimedia*, vol. 23, pp. 4272–4284, 2021.
- [38] M. Yang et al., "DPFL-Nets: Deep pyramid feature learning networks for multiscale change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6402–6416, Nov. 2022.
- [39] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4161–4170.
- [40] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8934–8943.
- [41] R. Xu, Z. Xiao, J. Huang, Y. Zhang, and Z. Xiong, "EDPN: Enhanced deep pyramid network for blurry image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 414–423.
- [42] X. Fu, B. Liang, Y. Huang, X. Ding, and J. Paisley, "Lightweight pyramid networks for image deraining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 6, pp. 1794–1807, Jun. 2020.
- [43] F. Wang et al., "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3156–3164.
- [44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [45] Y. Zhang et al., "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 286–301.
- [46] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
- [47] H. Lee, H.-E. Kim, and H. Nam, "SRM: A style-based recalibration module for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1854–1862.
- [48] D. Zhang, J. Shao, and H. T. Shen, "Kernel attention network for single image super-resolution," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 3, pp. 1–15, Jul. 2020.
- [49] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.
- [50] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 60–65.
- [51] 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.
- [52] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 2810–2818.
- [53] M. Singh, M. C. Govil, E. S. Pilli, and S. K. Vipparthi, "SOD-CED: Salient object detection for noisy images using convolution encoder-decoder," *IET Comput. Vis.*, vol. 13, no. 6, pp. 578–587, 2019.
- [54] T. Huang, S. Li, X. Jia, H. Lu, and J. Liu, "Neighbor2neighbor: Self-supervised denoising from single noisy images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14781–14790.
- [55] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "NAS-FPN: Learning scalable feature pyramid architecture for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7036–7045.
- [56] Y.-C. Chen, C. Gao, E. Robb, and J.-B. Huang, "NAS-DIP: Learning deep image prior with neural architecture search," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 442–459.
- [57] Z. Daquan, Q. Hou, Y. Chen, J. Feng, and S. Yan, "Rethinking bottleneck structure for efficient mobile network design," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland, 2020, pp. 680–697.
- [58] 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.
- [59] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
- [60] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
- [61] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1/4, pp. 259–268, 1992.
- [62] A. Paszke et al., "Automatic differentiation in pytorch," in *Proc. Workshop Autodiff*, 2017, pp. 1–4.

- [63] Y.-S. Chen, Y.-C. Wang, M.-H. Kao, and Y.-Y. Chuang, "Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6306–6314.
- [64] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 97–104.
- [65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.
- [66] 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.
- [67] Q. Zhang, Y. Nie, and W.-S. Zheng, "Dual illumination estimation for robust exposure correction," *Comput. Graph. Forum*, vol. 38, no. 7, 2019, pp. 243–252.
- [68] B. Cai et al., "A joint intrinsic-extrinsic prior model for retinex," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4000–4009.
- [69] 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.
- [70] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1185–1198, May 2011.
- [71] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
- [72] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
- [73] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. IEEE 19th Int. Conf. Image Process.*, 2012, pp. 1477–1480.
- [74] Q. Wang et al., "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.
- [75] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13713–13722.



Weiming Dong is a Member of the ACM.

Weiming Dong (Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from Tsinghua University, Beijing, China, in 2001 and 2004, respectively, and the Ph.D. degree in computer science from the University of Lorraine, France, in 2007. He is currently a Professor with the Sino-European Lab in Computer Science, Automation and Applied Mathematics and National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing. His research interests include image synthesis and image recognition.



Xingjia Pan received the B.S. degree in automation and finance from Nankai University, Tianjin, China, in 2015, and the Ph.D. degree in pattern recognition and intelligent system from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2020. He is currently a Senior Researcher with YouTu Lab, Tencent. His research interests include computer vision and machine learning.



Shideng Lin received the B.Sc. degree in vehicle engineering from the Nanjing University of Science And Technology, Nanjing, China, in 2017. He is currently working toward the M.Sc. degree with the University of Chinese Academy of Sciences, Beijing, China. His research interests include image and video processing and enhancement.



Fan Tang received the B.Sc. degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2019. He is currently an Associate Professor with the Institute of Computing Technology Chinese Academy of Sciences. His research interests include image synthesis and image recognition.



Changsheng Xu (Fellow, IEEE) is currently a Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and the Executive Director of the China-Singapore Institute of Digital Media. He was the recipient of the Best Associate Editor Award of *ACM Transactions on Multimedia Computing, Communications and Applications* in 2012 and the Best Editorial Member Award of *ACM/Springer Multimedia Systems Journal* in 2008. He was a Program Chair of ACM Multimedia 2009. He was an Associate Editor, a Guest Editor, General Chair, Program Chair, Area/Track Chair, Special Session Organizer, Session Chair, and TPC Member for more than 20 IEEE and ACM prestigious multimedia journals, conferences, and workshops. He is an Associate Editor for *ACM Transactions on Multimedia Computing, Communications and Applications* and the Editor-in-Chief of *ACM/Springer Multimedia Systems Journal*. He is an IAPR Fellow and ACM Distinguished Scientist.