Luminance-Aware Pyramid Network for Low-Light Image Enhancement

Jiaqian Li, Juncheng Li, Faming Fang, Fang Li, and Guixu Zhang

***Abstract*—Low-light image enhancement based on deep convolutional neural networks (CNNs) has revealed prominent performance in recent years. However, it is still a challenging task since the underexposed regions and details are always imperceptible. Moreover, deep learning models are always accompanied by complex structures and enormous computational burden, which hinders their deployment on mobile devices. To remedy these issues, in this paper, we present a lightweight and efficient Luminance-aware Pyramid Network (LPNet) to reconstruct normal-light images in a coarse-to-fine strategy. The architecture is comprised of two coarse feature extraction branches and a luminance-aware refinement branch with an auxiliary subnet learning the luminance map of the input and target images. Besides, we propose a multi-scale contrast feature block (MSCFB) that involves channel split, channel shuffle strategies, and contrast attention mechanism. MSCFB is the essential component of our network, which achieves an excellent balance between image quality and model size. In this way, our method can not only brighten up low-light images with rich details and high contrast but also significantly ameliorate the execution speed. Extensive experiments demonstrate that our LPNet outperforms state-of-the-art methods both qualitatively and quantitatively.**

***Index Terms*—Low-light image enhancement, luminance-aware guidance, multi-scale contrast feature, pyramid structure.**

I. Introduction

IMAGE with low illumination often suffers from severe degradations like low contrast, unexpected noise, and absence of natural colors due to the equipment constraints and inappropriate configurations. These drawbacks not only result in unpleasant visual effect, but also strongly affect high-level tasks such as image segmentation and object detection. Consequently, low-light image enhancement has great practical significance in computer vision. Currently, more and more researchers have expressed their keen interest in handling the problem of low-light image enhancement. Massive algorithms have been proposed to facilitate the subjective and objective quality of low-light images. Roughly speaking, there are three categories of the existing methods: histogram equalization based, Retinex-based, and learning-based approaches.

Histogram equalization (HE) [7] and its variants are the pioneering algorithms that enhance image contrast by expanding the dynamic range of pixels. However, these methods may lead to over-enhancement since the dependency between neighboring pixels is not considered. To overcome the disadvantages of global transformations based on HE, Wang *et al.* [8] proposed a variational way to determine a local transformation so that the histogram is redistributed locally and the brightness is preserved. Afterward, a generalized equalization model [9] integrating contrast enhancement and white balancing was established and showed favorable performance. Although these methods are relatively simple and fast, they are unfeasible to recover image details or colors.

Other traditional methods [1], [4], [10]–[12] are based on the Retinex theory [13] that decomposes an image into the pixel-wise product of reflectance and illumination. Following the common assumption, the reflectance component is often treated as an approximation of the enhanced image, so we only need to manipulate the estimated illumination map. Typically, two early heuristic algorithms SSR [11] and MSRCR [10] were devised to recover the illumination map assuming that there are some regularities in the colors of natural objects viewed under canonical illumination. However, these methods are prone to halo artifacts in the strong shadow transition area and the reflectance is in a limited range. To settle these issues, a naturalness preserved enhancement (NPE [1]) algorithm was employed for non-uniform illumination images. A fusion-based method MF [4] was proposed to blend multiple derivations of the initial illumination map, cooperated with different weights adapted to weak illumination circumstances. One limitation of this approach is that it always lacks authenticity in rich texture regions. Afterward, LIME [14] developed a structure-aware smoothing model to estimate the illumination map. However, the mentioned methods are challenging to apply for various scenarios due to the hand-crafted manipulations on the illumination map and exhaustive parameter tuning.

Manuscript received March 6, 2020; revised June 22, 2020 and August 13, 2020; accepted August 27, 2020. Date of publication September 3, 2020; date of current version September 24, 2021. This work was supported in part by the Key Project of the National Natural Science Foundation of China under Grant 61731009, in part by the National Natural Science Foundation of China under Grant 61871185 and 11671002, in part by the "Chenguang Program" supported by the Shanghai Education Development Foundation and Shanghai Municipal Education Commission under Grant 17CG25, and in part by the Fundamental Research Funds for the Central Universities, and Science and Technology Commission of Shanghai Municipality (No. 19 JC1420102, No. 18dz2271000). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chang-Su Kim. *(Jiaqian Li and Juncheng Li contributed equally to this work.) (Corresponding author: Faming Fang.)*

Jiaqian Li, Juncheng Li, Faming Fang, and Guixu Zhang are with the Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai 200062, China, and also with the School of Computer Science and Technology, East China Normal University, Shanghai 200062, China (e-mail: 51184506021@stu.ecnu.edu.cn; cvjunchengli@gmail.com; fmfang@cs.ecnu.edu.cn; gxzhang@cs.ecnu.edu.cn).

Fang Li is with the Shanghai Key Laboratory of PMMP, and School of Mathematical Sciences, East China Normal University, Shanghai 200241, China (e-mail: fli@math.ecnu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2020.3021243.

Digital Object Identifier 10.1109/TMM.2020.3021243

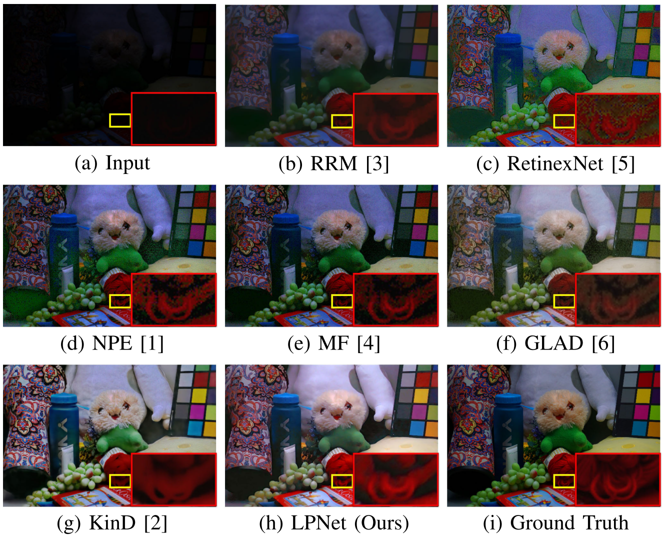


Fig. 1. An example of the enhanced images with state-of-the-art methods. The result generated by RRM [3] exists some dark regions and RetinexNet [5] is subjected to color distortion. The results in the second row (d)–(f) contain different degrees of noise. KinD [2] suffers from a little smooth as shown in (g). In contrast, the recovered image in (h) has rich details.

Recently, numerous learning-based methods have revealed their dominant performance in image denoising [15], [16] and image super-resolution [17], [18], as well as significantly facilitating the development of image enhancement [2], [5], [6], [19]–[23]. For instance, LLNet [19] proposed a deep autoencoder to enhance the image without over-amplifying. Combining the effectiveness of Retinex theory with convolutional neural networks, Wei *et al.* [5] attempted to embed the Retinex theory into CNN and devised the RetinexNet. Inspired by [5], KinD [2] introduced a novel structure comprising layer decomposition, reflectance restoration, and illumination adjustment to adjust light levels flexibly. In addition, bilateral grid learning was introduced by Gharbi *et al.* [22] for real-time evaluation. Similarly, DeepUPE [20] attempted to learn an image-to-illumination mapping and performed a bilateral grid in diverse lighting conditions. However, these methods do not particularly take edge information into account. By incorporating edge features, Ren *et al.* [21] proposed a hybrid network with a content stream and a salient edge stream to recover edge details when enhancing the degraded images.

Though these off-the-shelf approaches obtain comparable results in certain instances, they still have some drawbacks as shown in Fig. 1. Additionally, most of deep learning methods require sufficient computation resources and storage space, which are inadequate for mobile devices. In other words, there is still room for improvement in terms of performance and execution efficiency. To remedy these issues, we design a lightweight and efficient model for low-light image enhancement tasks. Specifically, we propose an innovative Luminance-aware Pyramid Network (LPNet), which consists of a luminance-aware refinement branch and two coarse feature extraction branches. To the best of our knowledge, it is the first coarse-to-fine architecture applied to image enhancement. Moreover, we elaborately devise a multi-scale contrast feature block (MSCFB), which is beneficial for feature representation and contrast information learning.

The main contributions of this paper are listed as follows:

i) We explore a coarse-to-fine architecture and devise a Luminance-aware Pyramid Network (LPNet) for low-light image enhancement. Extensive experiments demonstrate that our model can efficiently recover high-quality images with natural colors and rich details.

ii) A lightweight and effective block MSCFB is proposed as an essential component of our network, which can simultaneously extract image features at different scales and exploit contrast information.

iii) We propose a luminance-aware strategy to manipulate the illumination in the refinement branch progressively. Concretely, we introduce auxiliary guidance to learn the luminance between the input and target images gradually.

II. Related Work

Efficient Network Structure: Deep convolutional networks are always accompanied by numerous parameters, complex computation, and high demands for equipment, which makes them unfeasible to deploy on mobile devices. Lightweight structure design has gained increasing requirements and played an important role in recent years. Inception V2 [24] employed various small convolutional kernels to replace big kernels, which could lessen parameters and extract multi-scale features. ShuffleNet [25] utilized pointwise group convolution and channel shuffle that significantly alleviated the computational burden while maintaining high accuracy. ShuffleNet V2 [26] introduced channel shuffle and element-wise operations to speed up the model. MSRN [27] adopted convolutional kernels with distinct sizes to adaptively exploit the image features at different scales. Inspired by these methods, we propose a lightweight and effective multi-scale contrast feature block (MSCFB). Explicitly, MSCFB introduces channel split and shuffle strategies as well as distinct numbers of convolutional layers in each detached channel, which can explore multi-scale context information while averting unnecessary parameters.

Attention Mechanism: In recent years, a substantial number of attention modules have been proposed to emphasize informative features and suppress less valuable ones. For instance, Hu *et al.* [28] proposed a Squeeze-and-Excitation (SE) block that performed feature recalibration by modeling interdependencies between channels. Considering the positional relationship of each pixel is significant and cannot be neglected, NLNet [29] investigated a non-local operation to compute interactions between any two positions regardless of their distance. Later, GCNet [30] was proposed to simplify the NLNet utilizing a query-independent attention map for all positions to mitigate the computational complexity while preserving the accuracy. Recently, Fu *et al.* [31] devised a dual attention network with position and channel module in two branches to combine local features with global dependencies.

All of these methods are beneficial for high-level tasks such as object detection and scene segmentation, while they may have little effect on low-level tasks like image enhancement. To address this issue, Zheng *et al.* [32] first proposed the contrast-aware attention mechanism for single image super-resolution (SISR). However, SISR concentrates on the reconstruction of

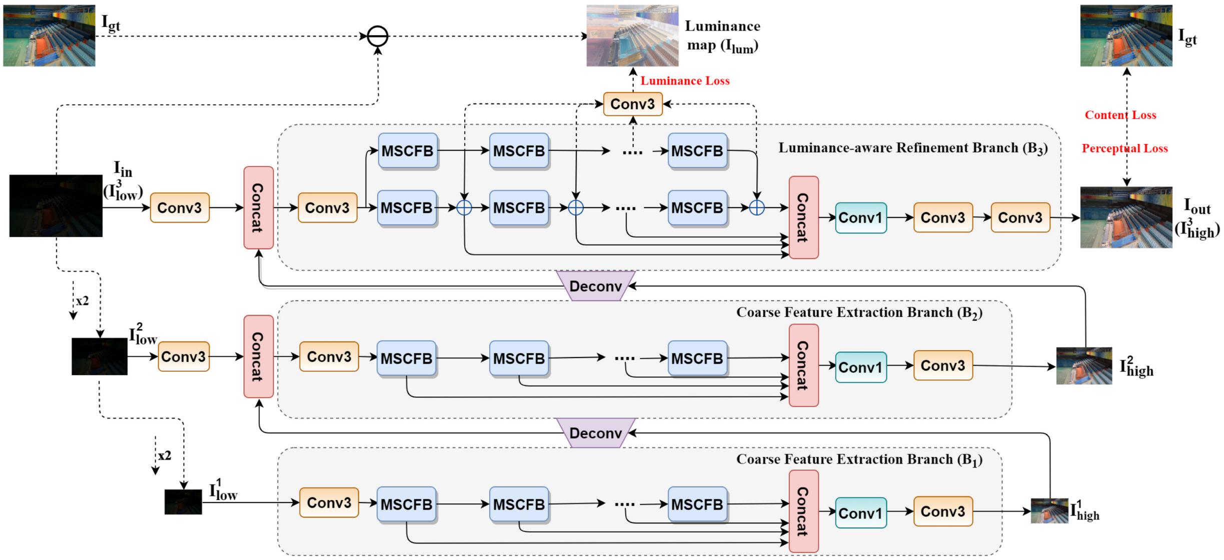


Fig. 2. The whole architecture of the proposed Luminance-aware Pyramid Network (LPNet). A coarse-to-fine framework that consists of two coarse feature extraction branches (B_1, B_2) and a luminance-aware refinement branch B_3.

high-frequency details rather than the improvement of contrast. Accordingly, image enhancement, aiming at improving the visual effect of the image, highly emphasizes the contrast information. Inspired by this idea, we introduce the contrast attention module to MSCFB, which is the first attempt for low-light image enhancement. Therefore, our proposed MSCFB can adaptively explore the contrast information and recover the hidden details in the dark regions.

Coarse-to-fine Architecture: The coarse-to-fine architecture has been widely used in image processing. Emily *et al.* [33] proposed a cascade Laplacian pyramid generative adversarial network breaking the generation into four levels for gradual refinements. In [34], a novel cascaded deep auto-encoder networks (CDAN) approach was utilized for face alignment, which achieved superior alignment accuracy with real-time speed. For dynamic scene deblurring, Seung *et al.* [35] imitated the conventional coarse-to-fine optimization and presented a multi-scale architecture to remove complicated motion blurs without estimating blur kernels. Fu *et al.* [36] introduced the mature Gaussian Laplacian image pyramid decomposition to the neural network for image deraining with a low parameter count. Similarly, Ren *et al.* [37] adopted a multi-scale approach to better train the proposed gated fusion network for preventing halo artifacts. All these methods favorably illustrate the effectiveness of the coarse-to-fine strategy, which preserves fine-grained detail information as well as a long-range dependency from coarser scales. Given the above insight, we employ a pyramid architecture in the low-light image enhancement task for the first time.

III. Luminance-Aware Pyramid Network

In this section, we describe the proposed Luminance-aware Pyramid Network (LPNet) in detail. As shown in Fig. 2, LPNet is essentially a coarse-to-fine framework, which consists of two coarse feature extraction branches (B_1, B_2) and a luminance-aware refinement branch B_3. In the coarse feature extraction branches, we utilize several multi-scale contrast feature blocks (MSCFBs) to extract global features at different scales. Since the input images of B_1 and B_2 obtain a lower resolution after downsampling, that is, a larger receptive field can be achieved. Next, these features are incorporated into the upper branch to exploit finer features. Finally, the extracted global image features from B_2 are delivered to the luminance-aware refinement branch B_3 for image enhancement. Similarly, we introduce several MSCFBs in B_3 to adaptively explore the contrast information and recover the hidden local details in the low-light images. Additionally, we propose a luminance-aware mechanism in B_3 for brightness adjustment by learning the luminance mapping between the input and target images progressively.

Define I_{in} and I_{out} as the input and output of our LPNet, respectively. I^i_{low} and I^i_{high} are the corresponding input and output of B_i branch where $i \in \{1, 2, 3\}$. To begin with, we gradually carry out downsampling operation with a factor of 2 to I_{in} for obtaining the I^i_{low}, expressed as

$$I^i_{low} = I^{i+1}_{low} \downarrow, i = 1, 2, \tag{1}$$

where $\downarrow$ is the average pooling operation and $I^3_{low} = I_{in}$. After receiving all levels of input images, we deliver them to the corresponding branch for feature extraction. It is worth mentioning that the output of the previous branch is upsampled to the identical size as the current input image. Then we concatenate them together in the current branch for more refined feature extraction, i.e.,

$$I^i_{high} = F_{B_i}([F_{conv3}(I^i_{low}), I^{i-1}_{high} \uparrow]), i = 2, 3, \tag{2}$$

where $F_{B_i}(\cdot)$ denotes the corresponding operation of B_i and $I^1_{high} = F_{B_1}(I^1_{low})$. F_{conv3} is a 3×3 convolutional layer, $\uparrow$ is a deconvolutional layer, and $[\cdot]$ is the concatenation operation. Among them, $I^3_{high} = I_{out}$ represents the enhanced image.

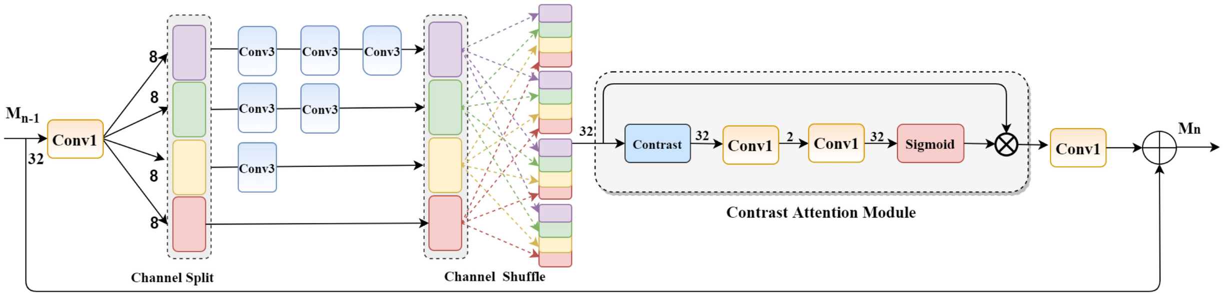


Fig. 3. The structure of our proposed multi-scale contrast feature block (MSCFB), which is the essential component of the LPNet. Where 32, 8, and 2 all denote the output channels of the corresponding convolutional layer.

During the training process, we propose a luminance-aware loss function, which consists of content loss, luminance loss, and perceptual loss. Therefore, given a collection of M image pairs $\{I_{in}^m, I_{gt}^m\}_{m=1}^M$, we aim to solve the following problem

$$\hat{\theta} = arg \min_{\theta} \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}_{total}(F_{LPNet_\theta}(I_{in}^m), I_{gt}^m), \tag{3}$$

where I_{in}^m and I_{gt}^m denote the input image and ground truth, respectively, θ is the parameter set and $F_{LPNet}(\cdot)$ represents our LPNet. Here, $\mathcal{L}_{total}(\cdot)$ is the luminance-aware loss function adopted to minimize the difference between the enhanced images and ground truth. Each module of the network will be described in the following part, and the definition of the luminance-aware loss will be introduced in Section III-D.

A. Multi-Scale Contrast Feature Block (MSCFB)

The receptive field plays an important role in the network design. A large receptive field not only provides rich context information but also learns the long-range relationship between pixels. Dilated convolutions [38] and pooling are two representative operations to expand the receptive field. However, the former will result in gridding effects and irrelevant long-ranged information. The latter will lack crucial details during the dimensionality reduction process. Based on the preceding insight, we propose an efficient module named multi-scale contrast feature block (MSCFB), which can extract different scales of image features and exploit contrast information. Besides, it is an independent module that can be used as a drop-in replacement in various networks flexibly.

Multi-scale Feature Extraction: Inspired by MSRN [27], we are committed to exploring multi-scale features for image enhancement. However, directly employing it will bring numerous unnecessary parameters and substantial inference cost. To lessen the model size and calculation burden, we introduce channel split and channel shuffle [25] into MSCFB.

As shown in Fig. 3, it initially goes through a 1 × 1 convolutional layer. Next, we employ channel split operation on the preceding feature maps, which are partitioned into four groups with one-quarter channels. Different from [27], we utilize distinct numbers of 3 × 3 convolutional layers in each group to extract multi-scale context information. Notice that two 3 × 3 kernels have the identical receptive field with a 5 × 5 kernel. Compared with diametrically applying a 5 × 5 or 7 × 7 convolutional kernel, this mode is capable of reducing the parameters while preserving a larger receptive field. Nevertheless, the outputs from a fixed group are barely related to the inputs within this group, which hinders information flow between each group and weakens feature representation. To facilitate the feature exchange and cross-scale communications between groups, the channel shuffle strategy is applied after the concatenation operation. According to the ShuffleNet [25], the channel shuffle operation is especially critical to the information routing among different channels. Furthermore, it contributes to enhanced learning capability without additional parameters, which is in accordance with the design principle of a lightweight module.

In general, our block achieves an excellent balance between the computational burden and feature representation.

Contrast Attention Module: Low-light image enhancement aims to eliminate the darkness and make the hidden contents visible. It is difficult to obtain texture details from a low-light image diametrically using off-the-shelf feature extraction blocks, since they may result in blur and noise in the reconstructed image. To remedy this problem, we suggest evaluating the contrast of the feature maps during feature extraction. It is known that the standard deviation reflects the dispersion degree of the pixel values from the mean in a clear image. The larger the standard deviation is, the higher the contrast of the image, so is the image quality. Consequently, we further introduce the contrast attention mechanism [32] into MSCFB to promote the exploited multi-scale information, which is beneficial for low-level vision tasks and dramatically facilitates performance improvement. As depicted inside the dotted box in Fig. 3, we first implement the standard deviation based on the global average pooling.

Denote $X = [x_1, x_2, \ldots, x_C]$, $X \in \Omega^{C\times H\times W}$ as a feature map. We use the standard deviation σ_c of the c-th element to evaluate the contrast degree, i.e.,

$$\sigma_c = \left(\frac{1}{HW} \sum_{(i,j)\in\Omega} (x_c^{i,j} - mean(x_c))^2 r \right)^{\frac{1}{2}}, \tag{4}$$

where $mean(x_c)$ is the mean of x_c, (i,j) denotes the position in the feature map. Then, we utilize two 1 × 1 convolutional layers that act as channel downscaling with a reduction ratio 16 to aggregate the contrast features. Each convolutional layer is accompanied by a ReLU for activation. Afterward, a sigmoid function is applied to calibrate the weight of each channel according to the importance of candidate features. Moreover, all the calculated weights are combined with the input features in element-wise multiplication. At the tail of MSCFB, a 1×1 convolutional layer is implemented for feature integration. Ultimately, we introduce local residual learning into our MSCFB to further improve the information flow.

Even though the contrast attention is similar to the common channel attention [28] to some extent, their design intentions are different. We conduct an interrelated ablation study of two attention modules in Section V. In general, by taking advantage of the contrast attention mechanism, our MSCFB can adaptively learn the contrast information and reconstruct the enhanced image with more high-frequency details.

B. Luminance-Aware Strategy

Currently, most low-light image enhancement algorithms directly learn the mapping between the input and target images. However, these methods are difficult to accurately capture the change of luminance, which often leads to overexposure or insufficient brightness in the reconstructed image. Inspired by the residual learning strategy presented in VDSR [39], we propose a simple but effective luminance-aware mechanism. To begin with, we calculate the difference between the input and target images and define it as the luminance map. Next, we construct a sub-network to manipulate the illumination and utilize it to guide the reconstruction process progressively.

As shown in Fig. 2, the top branch (B_3) is our proposed luminance-aware refinement branch, which can be divided into two parallel components. In the top part, we adopt N MSCFBs to constitute the luminance-aware mechanism for brightness adjustment. Meanwhile, identical number of MSCFBs are utilized to build up the bottom feature refinement part. As depicted in B_3, the feature maps are respectively delivered to two parts for different purposes after two convolutional layers.

Define the input of the first top and bottom MSCFB as L_0 and R_0 ($L_0 = R_0$), respectively. The output of each MSCFB in the luminance branch can be written as

$$L_n = F^n_{LUM}(L_{n-1}),\ n = 1, 2, \ldots, N, \tag{5}$$

where $F^n_{LUM}(\cdot)$ denotes the operation of the $n\text{-}th$ MSCFB in the luminance part. Similarly, the output of each MSCFB in the refinement branch can be defined as

$$R_n = F^n_{REF}(R_{n-1}),\ n = 1, 2, \ldots, N, \tag{6}$$

where $F^n_{REF}(\cdot)$ is the operation of the $n\text{-}th$ MSCFB in the refinement branch. For luminance awareness, we utilize each intermediate output from L_n to progressively guide the refinement stage by adding the corresponding R_n in an element-wise way. So we modify Eq.(6) as

$$R_n = F^n_{REF}(R_{n-1}) + F^n_{LUM}(L_{n-1}),\ n = 1, 2, \ldots, N. \tag{7}$$

To take advantage of image features, we adopt a hierarchical information distillation strategy. Since these features contain redundant information and significantly scale up the computational complexity, we first concatenate all the feature maps and utilize a 1×1 convolutional layer to aggregate them. Through this scheme, we can maintain the integrity of hierarchical features with fewer parameters. Finally, two 3×3 convolutional layers are applied to acquire the final RGB image.

In conclusion, based on the guidance of the luminance-aware mechanism, our LPNet can reconstruct images with appropriate brightness distribution.

C. Pyramid Architecture

The aforementioned luminance-aware refinement branch is capable of generating an enhanced image with content and luminance consistency. However, due to the impact of information loss during the excessive convolutional process, the reconstructed image is still lack of texture details. Thus, we adopt a pyramid structure through multi-level learning in a coarse-to-fine strategy as shown in Fig. 2. Different from previous works [35], [36], [40], the proposed LPNet only compares the enhanced image obtained at the finest scale with the ground truth rather than calculating the loss for each scale.

Essentially, our LPNet is a dynamically scalable framework that contains i levels, where we can select an appropriate value according to the actual demands. In our experiment, we focus on constructing a lightweight model and set $i = 3$. Different from previous works that utilize the same architecture on each branch, our LPNet contains three distinct branches. Although B_1 and B_2 consist of N MSCFBs with the same structure, they have different receptive fields. Besides, B_3 is designed to exploit the local image features, while B_1 and B_2 are devised to capture the global features since they allow a larger receptive field to explore the whole patch. Moreover, information from the coarser level is delivered to the next branch for more refined feature reconstruction. After incorporating all the global features into the finest branch B_3, we utilize them to improve local features and generate the final enhanced image. It should be pointed out that these three branches (B_1, B_2, and B_3) in our network are connected sequentially, and the model is trained in an end-to-end manner.

Generally speaking, fine-grained detail information as well as a long-range dependency from coarser scales can be preserved by the pyramid structure.

D. Loss Function

To improve the image quality both qualitatively and quantitatively, we propose a luminance-aware loss function with the following three components: 1) L1 loss used to reconstruct image content; 2) Luminance loss adopted to learn the difference of brightness between the low-light and target images; 3) VGG loss utilized to enhance the perceptual quality.

Content Loss: We use MAE to measure the overall structural similarity instead of MSE, which can avoid getting stuck in a local minimum. The L1 loss is formulated as follows:

$$\mathcal{L}_{cont} = \frac{1}{M} \sum_{m=1}^{M} \left\| F_{LPNet}(I_{in}^{m}) - I_{gt}^{m} \right\|_1 , \tag{8}$$

where $F_{LPNet}(I_{in})$ is the reconstructed image, I_{in} and I_{gt} represent the input image and ground truth, respectively.

Luminance Loss: As described in Section III-B, we present a luminance-aware strategy to guide image reconstruction, so that a luminance loss is utilized to supervise the learning process. We calculate the difference between each hierarchical luminance map and the expected one, i.e.,

$$\mathcal{L}_{lum} = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \left\| F_{lum}^{n}(I_{in}^{m}) - I_{lum}^{m} \right\|_1 , \tag{9}$$

where $F_{lum}^{n}(I_{in}^{m})$ denotes the predicted luminance map of the n-th MSCFB and I_{lum}^{m} is the expected one.

Perceptual Loss: Inspired by [41], we introduce the VGG loss as the perceptual measurement to utilize image semantic information and facilitate the visual quality of the enhanced image. We use the Euclidean distance to calculate the difference between feature maps, formulated as below,

$$\mathcal{L}_{vgg} = \frac{1}{C_t H_t W_t} \left\| \phi_t(F_{LPNet}(I_{in})) - \phi_t(I_{gt}) \right\|_2^2 , \tag{10}$$

where $\phi_t(\cdot)$ is the activation feature map of the t-th convolutional layer in VGG19. Here, C_t, H_t, W_t represent the dimensions of the corresponding feature maps, respectively.

Total Loss: We define the total loss as a weighted sum of the above-mentioned three components, i.e.,

$$\mathcal{L}_{total} = \mathcal{L}_{cont} + \mathcal{L}_{lum} + \mathcal{L}_{vgg}. \tag{11}$$

In the experiment, we empirically set the weight of each component as 1. Moreover, our model is trained end-to-end with the total loss until it converges.

IV. Experiments

A. Datasets and Metrics

Datasets: In the experiment, we adopt LOL [5], MIT5K [42], and SID [43] as our training datasets. Among them, LOL consists of 500 low/normal-light image pairs captured in real scenes, which is the first dataset used for low-light image enhancement. We divide it into three sets, 450, 35, and 15 image pairs for training, validation, and testing, respectively. As for the MIT5K dataset, we use 4,500 images for training, with the remaining 500 images for validation and testing by Expert C following [20], [22]. SID dataset contains raw sensor images shot by Sony and Fujifilm in both indoor and outdoor scenes. In our study, we merely employ the Sony set including 2,697 raw short-exposure images and 231 long-exposure images. The same protocol in [43] is implemented to divide the data into training, validation, and testing sets. Considering that models trained on the raw domain cannot be applied to regular RGB images directly, we convert the dataset from raw data to png format in the rawpy python package following the preprocessing pipeline in [44]. To verify the robustness of our model, we train the network utilizing the noisy low-light images without any post-processing.

Additionally, we further employ several benchmark datasets, involving MEF [45], NPE [1], DICM [46], and VV [47] for test. Among them, MEF is composed of 17 image sequences with multiple exposure levels, and we select one of the poor-exposed images from each multi-exposure set for evaluation. Since all these images have no corresponding ground truth, they are often utilized to verify the generalization of different methods in real-world scenarios.

Evaluation Metrics: In order to quantitatively evaluate the performance of our network, we select two commonly used metrics (i.e., PSNR and SSIM) to measure the content and structural similarity between the enhanced images and ground truth. Generally speaking, higher PSNR and SSIM values indicate better results and authentic human perception. Besides, we further utilize Natural Image Quality Evaluator (NIQE) [48] to assess the image quality, where a lower value indicates better performance.

B. Implementation Details

Our network is implemented in the Pytorch framework and trained on NVIDIA RTX 2080Ti GPU with ADAM optimizer for 200 epochs. We respectively utilize the learning rate of 10^{-5} as the initialized value on the LOL dataset and 10^{-4} for the MIT5K and SID datasets with halving every 100 epochs. During training, we augment the training data with rotation, flipping horizontally and vertically to promote the generalization of the network. In addition, we randomly extract 16 patches, each with a size of 96×96 as inputs. In the final model, we allocate $N = 4$ and the channel of each feature map is set as 32 to construct a lightweight network. Specifically, no additional training tricks are employed.

C. Comparisons With State-of-the-Art Methods

We compare our model with various state-of-the-art methods, and all the results are assessed with PSNR and SSIM. To be specific, we compare RRM [3], NPE [1], MF [4], RetinexNet [5], GLAD [6], and KinD [2] on the LOL [5] dataset, White-Box [49], D & R [50], HDRNet [22], DPE [51], and DeepUPE [20] on the MIT5K [42] dataset. Meanwhile, LIME [14], RetinexNet [5], KinD [2], GLAD [6], and LTS [43] are compared on the SID [43] dataset. For all the methods, we either implement the codes provided by the authors or show the results that are public on the websites. Furthermore, all the deep learning algorithms are trained and tested with the recommended parameter settings and implementation details for a fair comparison.

Quantitative Comparisons: Table I reflects the quantitative comparisons with state-of-the-art methods on the LOL, MIT5K, and SID datasets. Among them, the final PSNR/SSIM results denote the average value of the corresponding test datasets and the best results are highlighted in bold. We can perceive that our LPNet achieves favorable performance on both LOL and MIT5K datasets and outperforms other methods by a large margin. When comparing to the SID dataset, we retrain all the deep learning algorithms with the converted RGB images after preprocessing the raw data for comparison fairness. In this work,

TABLE I
QUANTITATIVE COMPARISONS WITH STATE-OF-THE-ART IMAGE ENHANCEMENT METHODS ON THE LOL [5], MIT5K [42], AND SID [43] DATASETS. THE BEST RESULTS ARE **HIGHLIGHTED** IN BOLD OR IN RED

Datasets	Methods	PSNR/SSIM↑	Param (M)↓	FLOPs (G)↓	Time (s)↓
LOL	RRM [3]	13.88/0.658	-	-	0.8652
	NPE [1]	16.97/0.589	-	-	5.8106
	MF [4]	18.79/0.642	-	-	0.3128
	RetinexNet [5]	16.77/0.559	1.23	6.79	0.5217
	GLAD [6]	19.72/0.704	0.93	4.37	0.4679
	KinD [2]	20.87/0.802	8.49	7.44	0.6445
	LPNet (Ours)	**21.46/0.802**	0.15	0.77	0.0183
MIT5K	White-Box [49]	18.57/0.701	8.56	26.10	5.6723
	D & R [50]	20.97/0.841	35.66	31.07	1.0760
	HDRNet [22]	21.96/0.866	0.23	1.06	0.0102
	DPE [51]	22.15/0.850	6.67	15.36	0.5077
	DeepUPE [20]	23.04/0.893	0.75	3.46	0.1256
	LPNet (Ours)	**24.53/0.906**	0.15	0.77	0.0166
SID	LIME [14]	19.10/0.572	-	-	2.2919
	RetinexNet [5]	18.37/0.532	1.23	6.79	4.3068
	KinD [2]	21.54/0.582	8.49	7.44	5.9937
	GLAD [6]	25.50/0.654	0.93	4.37	1.0714
	LTS [43]	27.20/0.700	7.76	17.88	3.1695
	LPNet (Ours)	**27.22/0.701**	0.15	0.77	0.0206

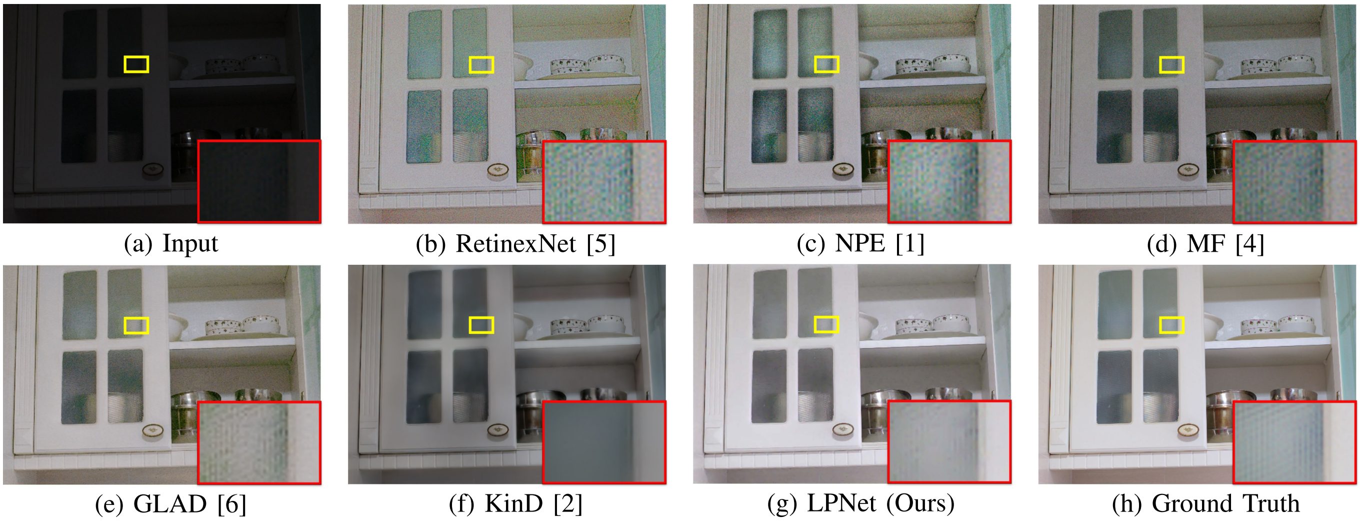
we mainly concentrate on the enhancement of low-light RGB images since raw data is usually unavailable due to the lack of expertise or unknown protocols. From Table I, we can observe that the average PSNR and SSIM achieved by our method are higher than other results, which sufficiently demonstrates the superiority and generalization ability of our approach.

Efficiency Investigation: In addition to the image quality, efficiency is a crucial indicator to measure the performance of an algorithm as well. Though massive deep learning methods have been proposed and achieved prominent performance, most of them are accompanied by excessive parameters or long execution time, which are unsuitable for small electronic devices. Due to the limited memory and resources of the equipment, it is necessary to design a lightweight and efficient model for image enhancement. Table I reports the performance of our method considering parameters, FLOPs (FLoating-point OPerations), and running time. Notice that the FLOPs are calculated with a patch size of 48×48. The running time is tested with the resolution of 600×400, 500×333, and 2128×1424 corresponding to the LOL, MIT5K, and SID datasets, respectively. It should be pointed out that RRM [3], NPE [1], MF [4], and LIME [14] are traditional algorithms implemented on Intel i5-9400 CPU. Meanwhile, other learning-based methods (including our LPNet) are tested on NVIDIA RTX 2080Ti GPU. Obviously, our model achieves better or similar performance over all state-of-the-art methods with few parameters and less running time. It demonstrates the LPNet is an efficient and lightweight model that achieves a balance among model size, execution time, and performance.

Subjective Evaluation: We present several visual comparisons against the aforementioned methods on three datasets. According to Fig. 4, we can clearly observe the enhanced image reconstructed by RetinexNet [5] suffers from different degrees of color distortion. While the recovered images produced by NPE [1] and MF [4] contain numerous noise that is especially noticeable in the zoomed areas. As shown inside the red rectangle of GLAD [6], there exists some noise in dark regions. Although the result of KinD [2] is comparable, it remains slightly dark and smooth, which leads to missing subtle details. In contrast, our LPNet can recover realistic colors and obtain more texture details without noise compared with other methods. Additionally, Fig. 5 demonstrates some representative comparisons with several competing approaches on the MIT5K dataset which includes a fraction of underexposed images. As depicted, though most methods can brighten up the input image, they still contain obstinate noise or exist color distortion due to unsatisfactory adjustments. For instance, the little girl's face from (b) to (f) is either underexposed or over-magnified while the result produced by our network looks more natural and conforms to human aesthetics. It indicates that our model plays a role of retouching and beautifying when it is run on a non-low-light image. We compare the visual quality of the results on the SID dataset in Fig. 6. It can be observed that although methods such as LIME [14] and RetinexNet [5] improve the illuminance to a certain extent, the enhancement results are subjected to color shift. KinD [2] suffers from relatively dark and smooth as shown in (d). The GLAD [6] algorithm can remove the unfavorable illuminance and promote the global contrast. However, the enhanced image still contains inevitable noise in some regions. The image generated by LTS [43] is comparable while it requires more resources and execution time. Compared to others, our method efficiently improves the image quality while preserving the natural colors and rich details.

All the results manifest that our method not only enhances the dark area without over-exposed artifacts, but also maintains the texture details with high contrast. In summary, with the impact of pyramid architecture, luminance-awareness guidance, and contrast attention mechanism, our LPNet can predict reasonable adjustments and reconstruct high-quality images.

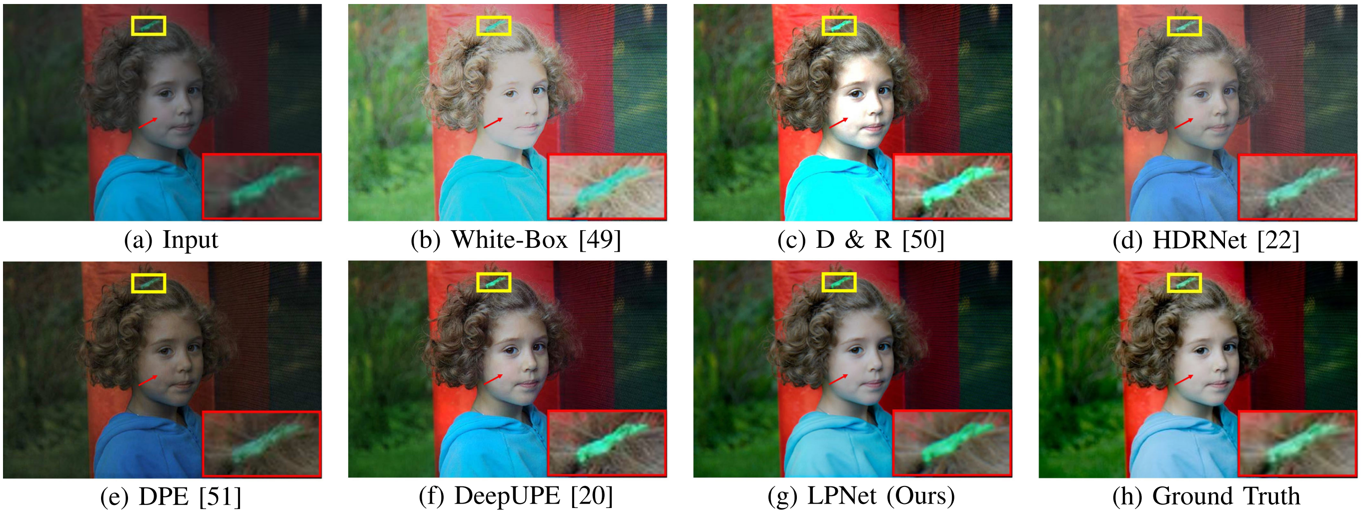
Generalization Ability: To further assess the robustness of our LPNet, we utilize the pre-trained model on the LOL to test some real-world low-light benchmarks including MEF [45], NPE [1], DICM [46], and VV [47]. We adopt blind image quality assessment NIQE [48] to appraise the performance with several representative methods for verifying the naturalness of the enhanced image. As presented in Table II, our LPNet shows its clear advantage against other methods on the whole. Specifically, LPNet outperforms all the competitors on the LOL, NPE, MEF, and DICM datasets. For the VV dataset, our results have marginal performance gaps between MF [4] and KinD [2]. This is a fairly remarkable demonstration of the relationship between quantified image authenticity and perceptual image quality. Furthermore, Fig. 7 shows a challenging case of a low-light image on the MEF dataset, where the original input is extremely dark with the imperceptible trace of texture in most regions. It is evident that the enhanced images from (b) to (g) contain some noise or artifacts, while our result has better visual effects. In other words, our method is robust for real-world extremely dark image enhancement.



(a) Input (b) RetinexNet [5] (c) NPE [1] (d) MF [4]

(e) GLAD [6] (f) KinD [2] (g) LPNet (Ours) (h) Ground Truth

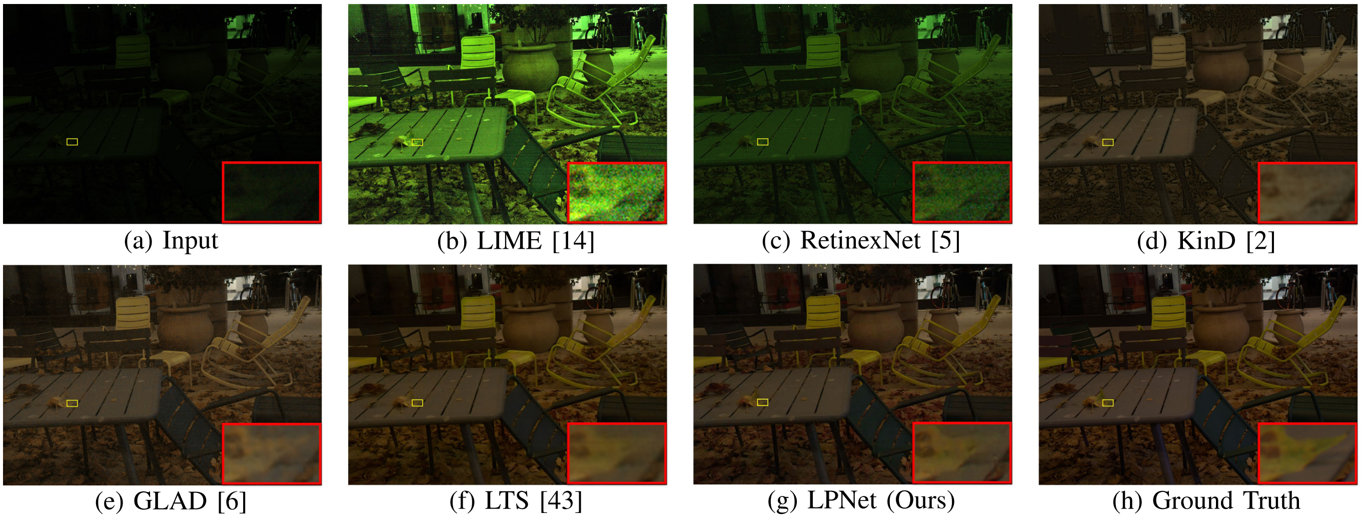
Fig. 4. Visual comparisons with state-of-the-art low-light image enhancement methods on the LOL [5] dataset.



(a) Input (b) White-Box [49] (c) D & R [50] (d) HDRNet [22]

(e) DPE [51] (f) DeepUPE [20] (g) LPNet (Ours) (h) Ground Truth

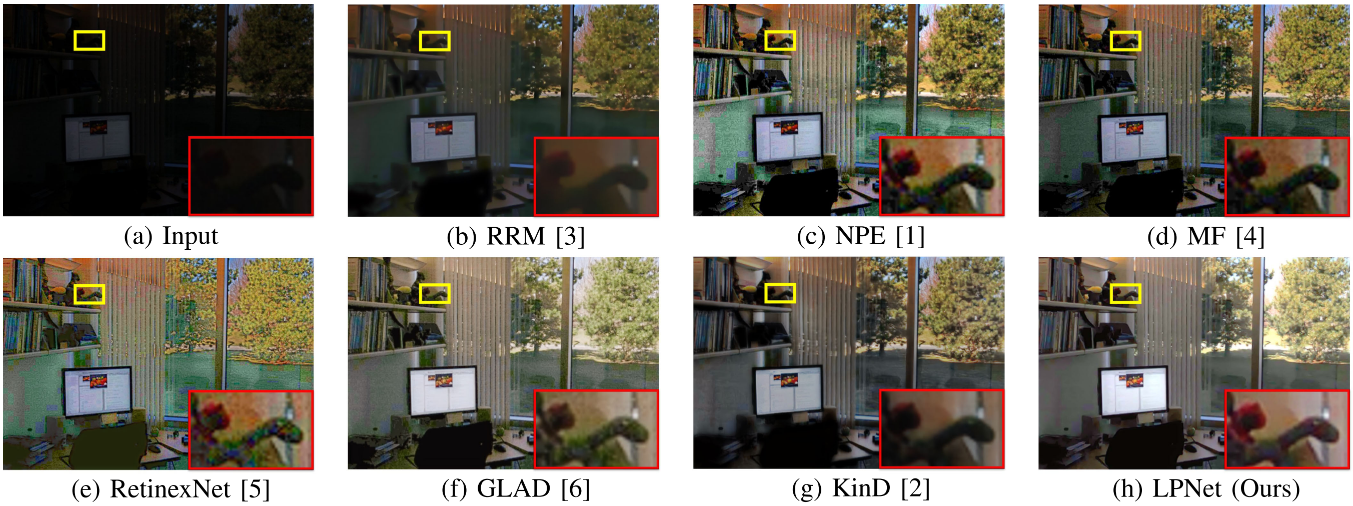
Fig. 5. Visual comparisons with state-of-the-art image enhancement methods on the MIT5K [42] dataset.



(a) Input (b) LIME [14] (c) RetinexNet [5] (d) KinD [2]

(e) GLAD [6] (f) LTS [43] (g) LPNet (Ours) (h) Ground Truth

Fig. 6. Visual comparisons with state-of-the-art low-light image enhancement methods on the SID [43] dataset.



(a) Input (b) RRM [3] (c) NPE [1] (d) MF [4]

(e) RetinexNet [5] (f) GLAD [6] (g) KinD [2] (h) LPNet (Ours)

Fig. 7. Visual comparisons with state-of-the-art low-light image enhancement methods on the MEF [45] dataset.

TABLE II
INVESTIGATIONS OF IMAGE NATURALNESS WITH NIQE METRIC ON SOME BENCHMARK DATASETS. THE BEST RESULT IS **HIGHLIGHTED** IN BOLD

Methods	Datasets				
	LOL [5]	NPE [1]	MEF [45]	DICM [46]	VV [47]
RRM [3]	5.8101	4.8452	4.1535	3.9560	3.8135
SRIE [52]	7.2869	3.9795	3.4577	3.7541	3.1376
LIME [14]	8.3777	4.2629	3.7159	3.7302	3.2104
NPE [1]	8.4390	3.9520	3.5378	3.8330	3.0313
MF [4]	8.8770	4.1096	3.4773	3.8207	**2.9297**
RetinexNet [5]	8.8785	4.5674	4.4755	4.5002	3.5248
GLAD [6]	6.4755	3.9699	3.3435	3.5704	3.0511
KinD [2]	5.1461	3.8826	3.3429	3.8608	2.9758
LPNet (Ours)	**4.5916**	**3.6173**	**3.3001**	**3.4958**	2.9977

TABLE III
ABLATION STUDY OF EACH COMPONENT IN OUR LPNET. √ AND × DENOTE WHETHER TO USE IT DURING TRAINING, RESPECTIVELY. WE TEST THE ENHANCED IMAGES WITH PSNR/SSIM METRICS

Case	Pyramid	Luminance	Contrast	Shuffle	Split	LOL	MIT5K	SID
1	×	√	√	√	√	21.36/0.789	24.32/0.904	26.75/0.685
2	√	×	√	√	√	20.16/0.788	24.18/0.903	26.26/0.680
3	√	√	×	√	√	18.86/0.783	22.11/0.886	26.77/0.689
4	√	√	√	×	√	21.20/0.801	24.40/0.905	27.10/0.696
5	√	√	√	×	×	20.61/0.797	23.89/0.900	26.33/0.680
6	√	√	√	√	√	**21.46/0.802**	**24.53/0.906**	**27.22/0.701**

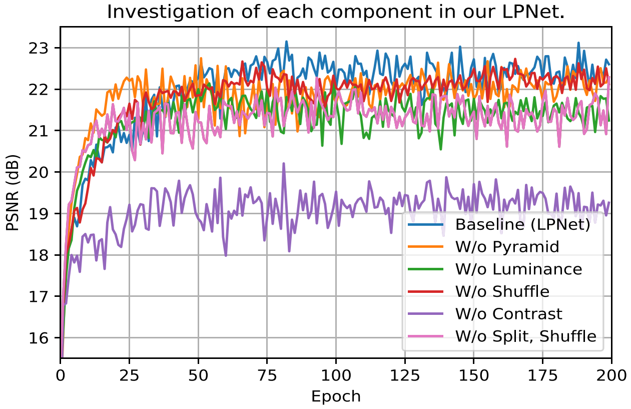


Fig. 8. Ablation study of each component in our LPNet training on the LOL dataset, where "W/o" denotes without the corresponding component.

V. ABLATION STUDY

In order to explore the effectiveness of the proposed method, we conduct a series of controlled experiments from different perspectives in this section. It is worth noting that we regard the result of our LPNet as the baseline.

A. *Effectiveness of the MSCFB Module*

MSCFB is the essential unit of our LPNet, which efficiently implements multi-scale feature extraction and contrast information learning. In this subsection, we provide the verification of the MSCFB module from two aspects: 1) evaluate the components within the MSCFB; 2) compare the MSCFB with several representative blocks.

1) Investigation of the Components within MSCFB: The MSCFB module is composed of channel split, channel shuffle, and contrast attention mechanism. In this part, we carry out a series of ablation studies to verify their necessity.

Study of Contrast Attention Module: In Table III, we show quantitative comparison results with (case 6) and without (case 3) the contrast attention module in MSCFB. It is evident that the model with high intensity contrast contributes to performance improvement of PSNR and SSIM on all the datasets. Meanwhile, we visualize the performance of each component during training in Fig. 8. Obviously, the purple curve is much lower than the blue curve during the entire training process. In other words, our network can steadily improve the performance with the assistance of the contrast attention module, especially low-level computer vision tasks, such as image enhancement and super-resolution.

Study of Channel Shuffle Operation: To validate the effectiveness of channel shuffle, we analyze the performance of removing it from MSCFB. As shown in Table III, the PSNR

TABLE IV
COMPARISONS WITH SEVERAL REPRESENTATIVE BLOCKS AND COMMON CHANNEL ATTENTION MODULE WITH PSNR/SSIM METRICS

Case	Module	LOL	MIT5K	SID
1	Residual Block	18.71 / 0.778	22.13 / 0.885	26.63 / 0.684
2	ShuffleNet Unit	18.53 / 0.769	21.98 / 0.882	26.48 / 0.682
3	Res2Net Module	18.70 / 0.774	21.98 / 0.884	26.67 / 0.683
4	Channel Attention	20.99 / 0.795	24.26 / 0.905	26.88 / 0.690
5	Baseline (LPNet)	**21.46 / 0.802**	**24.53 / 0.906**	**27.22 / 0.701**

and SSIM values of case 4 are slightly lower than that of case 6 on all the datasets. Correspondingly, it can be observed from the blue and red curves in Fig. 8, the performance of the model without channel shuffle operation is marginally inferior to the baseline. This is mainly because our network has a small number of feature channels, the difference of cases 4 and 6 is not palpable. With a large number of channels or groups, channel shuffle operation is more beneficial to promote the quality of enhanced image without additional parameters.

Study of Channel Split Scheme: In addition, we carry out an ablation experiment about removing both the channel split and shuffle schemes from our model. It is worth noting that directly discarding the split scheme will scale up a large number of parameters. In order to ensure a fair comparison, we utilize a 1×1 convolutional layer for dimension reduction in lieu of the split operation. From cases 5 and 6 in Table III, it is obvious that the PSNR and SSIM values degrade when we dismantle the split and shuffle operations in our LPNet. Besides, as shown in Fig. 8, the pink curve is generally below the blue and red curves during the steady stage of training. It demonstrates the superior capacity of the channel split scheme, which mitigates the number of parameters as well.

The preceding investigations illustrate the importance and efficiency of the proposed contrast attention mechanism, channel split, and shuffle operations in the MSCFB module. Consequently, the MSCFB can extract valuable features that contribute to enriching the texture information of the image.

2) Comparisons with Several Representative Blocks: In order to validate the effectiveness of our MSCFB, we compare it with the core feature extraction blocks proposed in ResNet [53], ShuffleNet [25], and Res2Net [54]. Moreover, we replace it with common channel attention to illustrate the superiority of contrast attention.

Compare with Feature Extraction Blocks: As we know, ResNet [53] has been widely used and revealed remarkable performance in various tasks. Considering the residual block is not a lightweight structure, we design a novel feature extraction block (MSCFB). Take the input channel of 32 as an example; a residual block has 18,432 parameters while our MSCFB merely contains 6,656 parameters. Besides, we make a comparison with two lightweight networks: ShuffleNet [25] and Res2Net [54]. To ensure an impartial experiment, we employ the same backbone as the infrastructure and only replace the block in the network. Furthermore, we make the total parameters of the models close by adjusting the number of channels or blocks. According to Table IV, we can perceive that our MSCFB achieves better results than other blocks with a smaller model size. In addition, as the performance curves displayed in Fig. 9, the results of other blocks are always below that of our MSCFB module. The main reason is that other blocks have no attention mechanism while our block cooperates with contrast attention, which can facilitate performance improvement. When comparing the results without contrast attention as shown in Table III, the proposed MSCFB is still comparable.

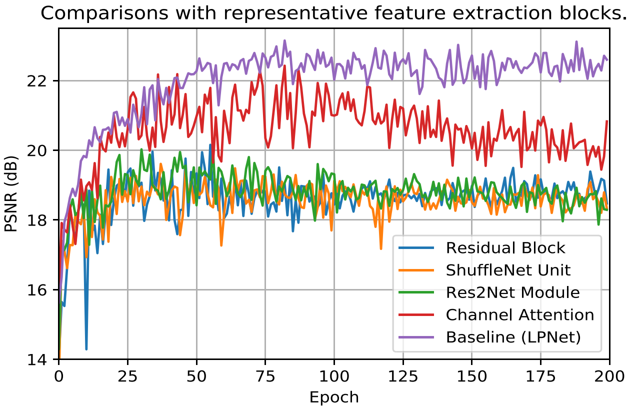


Fig. 9. Ablation study of the MSCFB module training on the LOL dataset.

Compare with Channel Attention Module: The idea of channel attention mechanism has been frequently employed in multiple previous works [28], [31], [55] and achieved tremendous performance improvements. We make a comparison with channel attention and adopt it to take the place of contrast attention module in our LPNet. As displayed in Table IV, the results of channel attention are inferior to that with the contrast attention mechanism. Accordingly, the performance of the red and purple curves in Fig. 9 is consistent with the results in the table. The main reason is that channel attention primarily concentrates on the activated high-value areas of a feature map, which is instructive to some high-level tasks such as classification, recognition, and detection. Conversely, low-level tasks not only focus on global information but also require local details to guide image reconstruction. Generally speaking, the contrast attention mechanism is more suitable for low-level computer vision tasks.

The above ablation studies and analyses indicate that the proposed MSCFB module achieves more superior performance with fewer parameters compared to other blocks.

B. Investigation of the Number of MSCFB

The number of MSCFB will significantly influence the number of parameters and performance of our LPNet. As shown in Fig. 10, we can observe that the result of our model performs worse when the number of MSCFB (N) decreases. However, the parameter quantity will increase as the block scales up. Meanwhile, the execution time will become longer when N gradually enlarges. In view of the trade-off between performance and efficiency, the number of MSCFBs should be selected according to the actual demand. It is worth noting that the performance gap becomes small between the results of $N = 5$ and $N = 4$.

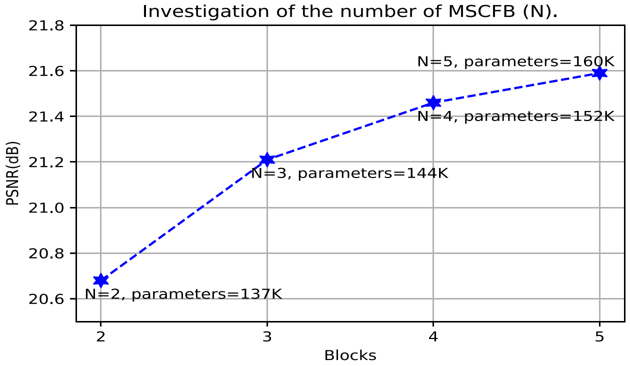


Fig. 10. Investigation of the number of MSCFB (N) on the LOL dataset, where "K" represents one thousand.

Therefore, in this paper, we set $N = 4$ for keeping in line with the design principle of a lightweight module, which can achieve favorable results with an acceptable number of parameters.

C. *Effectiveness of Network Architecture*

In this part, we verify the effectiveness of each component proposed in our LPNet on the LOL, MIT5K, and SID datasets, including the pyramid structure and luminance-aware strategy.

Study of Luminance-aware Strategy: We evaluate the validity of the luminance-aware mechanism by directly discarding it. At the same time, we remove the luminance loss in Eq.(11) for training. As shown in Table III, cases 2 and 6 denote the comparisons of these two methods. Obviously, the PSNR declines significantly once the luminance-aware part is removed. We find that the PSNR varies drastically on the LOL (1.30 dB) and SID (0.96 dB) datasets whereas it is not obvious on the MIT5K (0.35 dB) dataset. The main reason is that LOL and SID are low-light image datasets whose input images are extremely dark while the MIT5K dataset has a limited number of underexposed images. In other words, our luminance-aware guidance is specifically beneficial for low-light image enhancement. Accordingly, from the blue and green curves in Fig. 8, the PSNR value of the network without luminance-aware is below the baseline excluding the first a few epochs. This shows the effectiveness of our proposed strategy.

Study of Pyramid Structure: To assess the effectiveness of the pyramid structure, we build a new model termed as PNet, which removes the pyramid framework and only retains the top branch B_3. Meanwhile, the parameter quantities of PNet are at the same magnitude as the baseline by adjusting the number of channels for fairness. In Table III, cases 1 and 6 represent the PNet and LPNet, respectively. The PSNR and SSIM values on three datasets go down marginally once the pyramid structure is removed. In addition, as presented in Fig. 8, the orange curve is generally below the blue curve during the stable phase of training. Since the introduced pyramid structure can extract global image features that facilitate the improvement of local features, which significantly enriches the diversity of features. Therefore, our LPNet is able to reconstruct clear and accurate images compared to PNet.

TABLE V
ABLATION STUDY OF EACH COMPONENT IN THE LOSS FUNCTION. $\surd$ AND × DENOTE WHETHER TO USE IT DURING TRAINING, RESPECTIVELY. WE TEST THE ENHANCED IMAGES WITH PSNR/SSIM METRICS

Case	L_{cont}	L_{lum}	L_{vgg}	LOL	MIT5K	SID
1	$\surd$	×	×	20.37 / 0.783	24.02 / 0.899	26.81 / 0.699
2	$\surd$	$\surd$	×	21.06 / 0.789	24.34 / 0.903	27.00 / 0.699
3	$\surd$	×	$\surd$	20.84 / 0.798	24.22 / 0.905	27.02 / 0.694
4	$\surd$	$\surd$	$\surd$	**21.46 / 0.802**	**24.53 / 0.906**	**27.22 / 0.701**

In general, these analyses manifest that with the guidance of the luminance map, the enhanced image obtains appropriate brightness distribution. Moreover, our model can extract fine-grained image features and generate an enhanced image with rich texture details due to the ability of the pyramid structure.

D. *Effectiveness of Luminance-Aware Loss Function*

In this paper, a luminance-aware loss function is proposed for training the model. In Table V, we confirm the importance of each component in the loss function. It can be observed that the performance of training with the total loss is superior to that of removing any component from it. Accordingly, we present several visual comparisons in Fig. 11. It should be pointed out that L1 loss is used as the basic content loss as shown in (b). Then we successively add Luminance loss, VGG loss and both of them, which correspond to (c), (d), and (e), respectively. We can find out that the area inside the red rectangle in (b) is darker than the counterpart in (c). To solve this issue, we introduce luminance loss to avoid insufficient brightness. The result in (c) confirms its effectiveness while the patterns on the plates are still fuzzy and difficult to distinguish. Afterward, we employ VGG loss to promote the visual quality of the enhanced image. However, the plates at the lower-left corner in (d) are a little dark compared to the results optimized with the total loss in (e).

In summary, the above ablation studies sufficiently demonstrate each component of the loss function is indispensable. Furthermore, we can draw a conclusion that our proposed luminance-aware loss is effective to explore appropriate brightness and facilitate the image details.

E. *Investigation of High-Level Noise Image*

Extremely low-light imaging with limited illumination and short exposure is always subjected to high-level noise. To further verify the robustness of our model, we manually add white Gaussian noise with noise level $\sigma = 15, 30, 50$ on the MIT5K dataset for training and testing. As shown in Fig. 12, we can find our solution performs enhancing and denoising simultaneously. However, it should be pointed out that the recovered images (e) and (f) are smooth and lack some edge details due to the influence of noise. Especially, $\sigma = 50$ is a relatively high noise level which will severely destroy the image content, so that it is difficult to generate a clear and accurate image. In general, even though the presence of massive noise will disturb the training process and degrade the performance to some extent, our model is still effective. In the future, we hope to further analyze the impact of high-level noise on low-light image enhancement tasks.

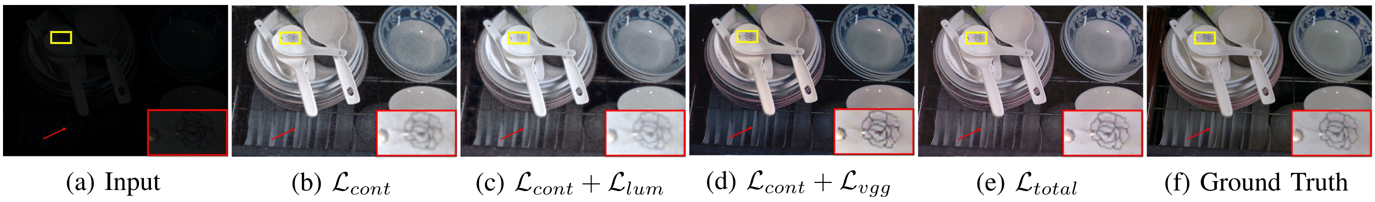

(a) Input
(b) $\mathcal{L}_{cont}$
(c) $\mathcal{L}_{cont} + \mathcal{L}_{lum}$
(d) $\mathcal{L}_{cont} + \mathcal{L}_{vgg}$
(e) $\mathcal{L}_{total}$
(f) Ground Truth

Fig. 11. Visual results of loss component ($\mathcal{L}_{cont}$, $\mathcal{L}_{lum}$, $\mathcal{L}_{vgg}$) in the Luminance-aware loss function on the LOL dataset.

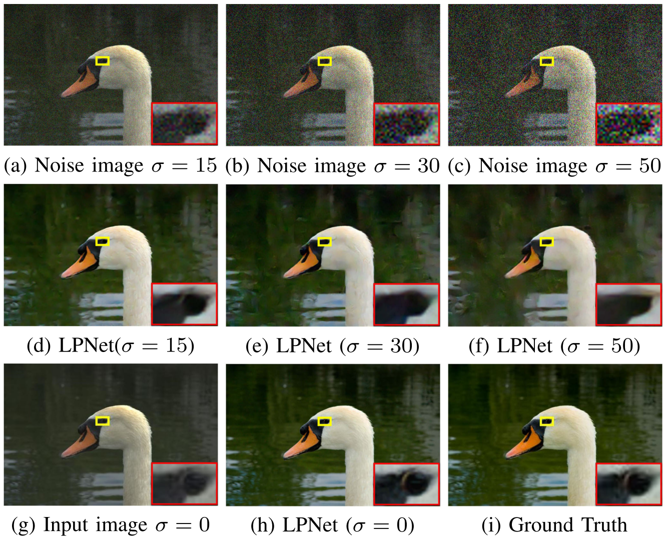

(a) Noise image $\sigma = 15$
(b) Noise image $\sigma = 30$
(c) Noise image $\sigma = 50$
(d) LPNet($\sigma = 15$)
(e) LPNet ($\sigma = 30$)
(f) LPNet ($\sigma = 50$)
(g) Input image $\sigma = 0$
(h) LPNet ($\sigma = 0$)
(i) Ground Truth

Fig. 12. Visual comparisons of the LPNet tested with different noise level images on the MIT5K dataset. The first row (a)–(c) are the simulated noise images based on the clear input image (g), (d)–(f) and (h) are the recovered images obtained by our method, and (i) represents the ground truth.

In addition, we strive to promote our method that can brighten up luminance while preserving rich details from the high-level noise image.

VI. Discussion and Limitation

Low-light image enhancement is a challenging but practical task, which is commonly used in various platforms such as smartphones, cameras, and embedded devices. Given the impracticality of exploiting large and deep networks on mobile equipment due to strict latency constraints, devising a lightweight and effective model is critical. Though our method achieves promising results in most cases with fewer parameters and faster speed, it still has some limitations. For instance, the result reconstructed by our LPNet in Fig. 7 (h) lacks some texture details hidden behind the reflective objects like glass or monitor. Additionally, it is unfeasible to recover the refined edge information under the excessive noise as shown in Fig. 12 (f). In future work, we aim to yield further improvements with semantic analysis to boost image quality. Moreover, we strive to build a large-scale paired dataset with diverse data distribution for low-light image enhancement of real scenes.

VII. Conclusion

In this paper, we remedy the low-light image enhancement problem by introducing an innovative Luminance-aware Pyramid Network (LPNet). The main idea is to construct a pyramid architecture across multi-level learning in a coarse-to-fine strategy, which consists of two coarse feature extraction branches and a luminance-aware refinement branch. Besides, a lightweight and efficient feature extraction block (MSCFB) is proposed to build up the entire framework, which strikes an excellent trade-off among performance, model size, and execution time. In addition, we employ a compound luminance-aware loss function that facilitates the visual quality of the enhanced image. Extensive experiments and ablation studies have illustrated our solution is efficacious both qualitatively and quantitatively, which has great potential to apply in low-light image/video enhancement tasks on mobile devices.

Acknowledgment

The authors would like to thank Dr. J. Qin from the University of Kentucky for her linguistic assistance during the preparation of this manuscript.

References

[1] X. Fu, D. Zeng, Y. Huang, X. Ding, and X.-P. Zhang, "A variational framework for single low light image enhancement using bright channel prior," in *Proc. IEEE Global Conf. Signal Inform. Process.*, 2013, pp. 1085–1088.

[2] 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.

[3] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-revealing low-light image enhancement via robust retinex model," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, Jun. 2018.

[4] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Process.*, vol. 129, pp. 82–96, 2016.

[5] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proc. Brit. Mach. Vis. Conf.*, 2018.

[6] 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 2018)*, 2018, pp. 751–755.

[7] S. M. Pizer *et al.*, "Adaptive histogram equalization and its variations," *Comput. Vis., Graphics, and Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.

[8] W. Wang and M. K. Ng, "A variational histogram equalization method for image contrast enhancement," *SIAM J. Imag. Sci.*, vol. 6, no. 3, pp. 1823–1849, 2013.

[9] H. Xu, G. Zhai, X. Wu, and X. Yang, "Generalized equalization model for image enhancement," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 68–82, Jan. 2013.

[10] D. J. Jobson, Zia-ur Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image processing*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

[11] D. J. Jobson, Zia-ur 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.

[12] S. Hao, X. Han, Y. Guo, X. Xu, and M. Wang, "Low-light image enhancement with semi-decoupled decomposition," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.2969790.

[13] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.

[14] 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. 2016.

[15] S. In Cho and S.-J. Kang, "Gradient prior-aided cnn denoiser with separable convolution-based optimization of feature dimension," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 484–493, Feb. 2018.

[16] Z. Jin, M. Z. Iqbal, D. Bobkov, W. Zou, X. Li, and E. Steinbach, "A flexible deep cnn framework for image restoration," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 1055–1068, Apr. 2020.

[17] X. Yang *et al.*, "Drfn: Deep recurrent fusion network for single-image super-resolution with large factors," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 328–337, Feb. 2018.

[18] Z. He *et al.*, "Mrfn: Multi-receptive-field network for fast and accurate single image super-resolution," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 1042–1054, Apr. 2020.

[19] 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.

[20] R. Wang, Q. Zhang, C.-Wi. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6849–6857.

[21] W. Ren *et al.*, "Low-light image enhancement via a deep hybrid network," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4364–4375, Sep. 2019.

[22] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Trans. Graph. (TOG)*, vol. 36, no. 4, p. 118, 2017.

[23] Y. Wang *et al.*, "Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement," pp. 2015–2023, 2019.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 448–456.

[25] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2018, pp. 6848–6856.

[26] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 116–131.

[27] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 517–532.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[29] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[30] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Oct. 2019, pp. 1971–1980.

[31] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3146–3154.

[32] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2024–2032.

[33] E. L. Denton, S. Chintala, R. Fergus *et al.*, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Advances Neural Inf. Process. Syst.*, pp. 1486–1494, 2015.

[34] R. Weng, J. Lu, Y.-P. Tan, and J. Zhou, "Learning cascaded deep auto-encoder networks for face alignment," *IEEE Trans. Multimedia*, vol. 18, no. 10, pp. 2066–2078, Oct. 2016.

[35] S. Nah, T. H. 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.

[36] X. Fu, B. Liang, Y. Huang, X. Ding, and J. Paisley, "Lightweight pyramid networks for image deraining," *IEEE Trans. Neural Netw. Learn. Syst.*, 2019.

[37] W. Ren *et al.*, "Gated fusion network for single image dehazing," pp. 3253–3261, 2018.

[38] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv:1511.07122*, 2015.

[39] 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.*, 2016, pp. 1646–1654.

[40] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8174–8182.

[41] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2016, pp. 694–711.

[42] 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.

[43] 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.

[44] C. Chen, Q. Chen, M. N. Do, and V. Koltun, "Seeing motion in the dark," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 3185–3194.

[45] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE TIP*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.

[46] C. Lee, C. Lee, and C. Kim, "Contrast enhancement based on layered difference representation of 2d histograms," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5372–5384, Dec. 2013.

[47] V. Vonikakis, I. Andreadis, and A. Gasteratos, "Fast centre-surround contrast modification," *IET Image Process.*, vol. 2, no. 1, pp. 19–34, 2008.

[48] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a completely blind image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2012.

[49] Y. Hu, H. He, C. Xu, B. Wang, and S. Lin, "Exposure: A white-box photo post-processing framework," *ACM Trans. Graph. (TOG)*, vol. 37, no. 2, p. 26, 2018.

[50] J. Park, J.-Y. Lee, D. Yoo, and I. S. Kweon, "Distort-and-recover: Color enhancement using deep reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5928–5936.

[51] 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.

[52] 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.*, 2016, pp. 2782–2790.

[53] 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.

[54] S. Gao *et al.*, "Res2net: A new multi-scale backbone architecture," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[55] Y. Zhang *et al.*, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.