Residual Quotient Learning for Zero-Reference Low-Light Image Enhancement

Chao Xie, *Member, IEEE*, Linfeng Fei, Huanjie Tao, Yaocong Hu, Wei Zhou, Jiun Tian Hoe, *Student Member, IEEE*, Weipeng Hu, and Yap-Peng Tan, *Fellow, IEEE*

***Abstract*— Recently, neural networks have become the dominant approach to low-light image enhancement (LLIE), with at least one-third of them adopting a Retinex-related architecture. However, through in-depth analysis, we contend that this most widely accepted LLIE structure is suboptimal, particularly when addressing the non-uniform illumination commonly observed in natural images. In this paper, we present a novel variant learning framework, termed residual quotient learning, to substantially alleviate this issue. Instead of following the existing Retinex-related decomposition-enhancement-reconstruction process, our basic idea is to explicitly reformulate the light enhancement task as adaptively predicting the latent quotient with reference to the original low-light input using a residual learning fashion. By leveraging the proposed residual quotient learning, we develop a lightweight yet effective network called ResQ-Net. This network features enhanced non-uniform illumination modeling capabilities, making it more suitable for real-world LLIE tasks. Moreover, due to its well-designed structure and reference-free loss function, ResQ-Net is flexible in training as it allows for zero-reference optimization, which further enhances the generalization and adaptability of our entire framework. Extensive experiments on various benchmark datasets demonstrate the merits and effectiveness of the proposed residual quotient learning, and our trained ResQ-Net outperforms state-of-the-art methods both qualitatively and quantitatively. Furthermore, a practical application in dark face detection is explored, and the preliminary results confirm the potential and feasibility of our method in real-world scenarios.**

***Index Terms*— Low-light image enhancement, residual quotient learning, zero reference, deep learning.**

Received 17 June 2024; revised 20 October 2024; accepted 2 December 2024. Date of publication 24 December 2024; date of current version 13 January 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62203012, Grant 62102320, and Grant 61901221; in part by the Natural Science Foundation of the Anhui Higher Education Institutions of China under Grant 2023AH030020; in part by the State Visiting Scholar Program of China Scholarship Council under Grant 202208320239; and in part by the National College Student Practice and Innovation Training Program under Grant 202410298020Z. The associate editor coordinating the review of this article and approving it for publication was Dr. Shiqi Wang. *(Corresponding author: Chao Xie.)*

Chao Xie and Linfeng Fei are with the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China (e-mail: chaoxie@njfu.edu.cn).

Huanjie Tao is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China.

Yaocong Hu is with the School of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China.

Wei Zhou is with the School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, Guangdong 510006, China.

Jiun Tian Hoe, Weipeng Hu, and Yap-Peng Tan are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798.

Digital Object Identifier 10.1109/TIP.2024.3519997

I. Introduction

LOW-LIGHT image enhancement (LLIE) is a fundamental and long-standing task in the field of image processing with a wide range of practical applications [1], [2], [3], such as face detection, video surveillance, and autonomous driving, to name a few. LLIE aims to improve the visibility and visual perception of low-light images, typically captured under poor lighting conditions, by increasing brightness, enhancing contrast, restoring color, and suppressing noise. This technique is of great significance, since it serves as a critical preprocessing procedure for many high-level computer vision tasks, particularly in nighttime or other poorly lit scenes.

Despite great challenges posed by LLIE, various methods have been developed and proposed over the past decades. Generally, these LLIE methods can be classified into two main categories: 1) conventional methods and 2) deep-learning-based methods.

Initial attempts [4], [5], [6], [7], [8], [9], [10] mostly belong to the former. Among them, a branch of Retinex-based methods has attracted relatively more attention. According to Retinex theory [11], [12], a low-light image can be decomposed into its reflectance and illumination components. Further considering the ill-posedness of this decomposition, different image priors have been additionally introduced into the respective optimization models to further regularize the problem and obtain a stable solution. Consequently, the light enhancement task can be accomplished by optimizing the components iteratively. Even though these methods seem promising in certain cases, they rely heavily on handcrafted priors and manual parameter tuning, which often makes them inadequate for real-world applications [13].

Owing to the spectacular success of deep learning [14], the latter category has recently emerged as the mainstream approach to LLIE. This category, namely deep-learning-based methods, is typically characterized and distinguished by the extensive use of various neural networks. For better analysis, we further divide it into two groups based on the specific network architecture used (see Figs. 1(a) and 1(b)). As a starting point, Lore et al. [15] proposed a variant of the stacked sparse denoising autoencoder named LLNet for simultaneous image brightness enhancement and denoising. Later, considering multi-scale feature extraction, Lv et al. [16] presented a multi-branch low-light enhancement network (MBLLEN), which demonstrated improved performance due to its enriched representation ability. Subsequently, more useful deep learning

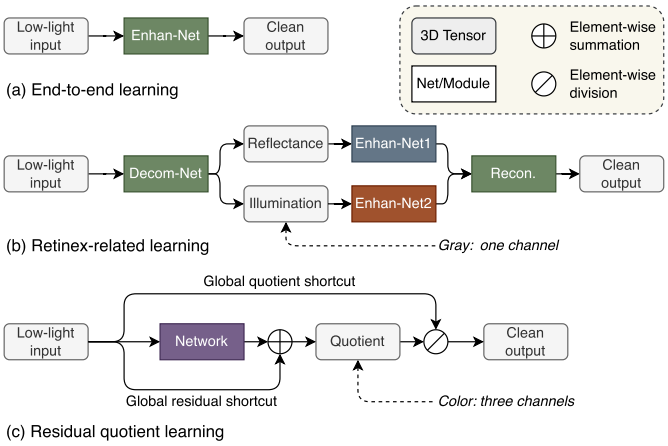


Fig. 1. Comparison of existing learning frameworks with ours for low-light image enhancement. (a) End-to-end learning framework, accepted by LLNet, MBLLEN, and DLN. (b) Retinex-related learning framework, accepted by Retinex-Net, KinD, RRDNET, URetinex-Net, and PairLIE. (c) Our residual quotient learning framework. The core insight and distinction of our framework is that it lays particular stress on the adaptive mapping of the underlying quotient, which physically represents illumination but with full color channels preserved. Zoom in for the best view.

techniques have been incorporated into the basic end-to-end learning framework to make further progress. For example, Ren et al. [17] designed a hybrid two-stream network to simultaneously learn the global content and salient structures of images. Wang et al. [18] utilized residual learning [19] to ease potential training difficulties. The pyramid network and multi-level Laplacian pyramid are introduced by LPNet [20] and DSLR [21], respectively. These end-to-end methods contribute to integrating feature representations effectively and efficiently for LLIE.

Instead of building LLIE networks using a complete end-to-end fashion, many researchers have begun redesigning their models under the guidance of Retinex theory [11], [12]. As a result, the Retinex-related learning framework has come into being and gradually become by far the most prevalent and dominant LLIE approach. Its basic concept is to first decompose the low-light input into specific components (usually reflectance and illumination), and then enhance each component individually. The entire process is carried out in a learning manner with the collaboration of several specifically constructed subnets. Therefore, it can be essentially regarded as an appropriate combination of conventional Retinex theory and modern neural networks. In addition, according to the recent comprehensive survey conducted in [22], about 35% of the deep-learning-based LLIE methods have accepted this Retinex-related architecture, further confirming its popularity. Although the Retinex-related learning framework generally achieves better performance than the end-to-end counterpart, we find it still suboptimal for the following reasons: 1) The structure is relatively overcomplicated and redundant compared to the end-to-end one. 2) The necessity and superiority of internal cooperation among the subnets remain questionable and unproven. 3) The whole framework is inadequate for addressing complex non-uniform illumination.

Therefore, in this paper, we present a novel variant learning framework for LLIE, and name it *residual quotient learning*, as depicted in Fig. 1(c). Specifically, in our framework, the entire LLIE mapping is recast into a straightforward two-step learning task. The first step involves a residual learning problem. Concretely, we take the low-light image as input and feed it to a trainable network to predict the residual between the underlying illumination and the low-light image itself, so the unknown illumination can be calculated by attaching a global residual shortcut. Afterward, we alternatively view the estimated illumination as the quotient of the low-light observation divided by the desired recovery. Hence, the final clean output can be readily generated through the second step (*i.e.*, a trivial element-wise division operation), which is similarly realized by appending another global quotient learning shortcut.

Compared to current Retinex-related and end-to-end architecture, our residual quotient learning framework eliminates the redundancy of the former and retains the simplicity of the latter, while still being physically Retinex-explainable. From the above analysis, it can be inferred that our presented framework is intuitively simpler and more effective as it simultaneously combines the advantages of both structures. Therefore, our learning framework will facilitate the subsequent network training and even improve the overall performance.

Based on this learning framework, we propose a lightweight yet effective network ResQ-Net for practical LLIE. Compared to previous work, our main contributions can be summarized as follows:

- In contrast to most existing Retinex-related schemes that adopt a decomposition-enhancement-reconstruction process, we present a novel residual quotient learning framework in which the primary LLIE task is explicitly recast as adaptively estimating the latent quotient. To the best of our knowledge, our work is the first successful attempt to apply this type of learning framework to LLIE, achieving encouraging results.
- By virtue of our presented learning framework, we propose a residual quotient net (ResQ-Net) for practical LLIE. Owing to its elaborately designed structure and well-formulated loss function, our ResQ-Net offers flexible training capabilities, allowing for zero-reference optimization. This substantially enhances the generalization and adaptability of our whole system.
- Extensive benchmark evaluations are conducted to verify the validity of the above two main contributions. Our final trained ResQ-Net is demonstrated to surpass recent state-of-the-art methods both quantitatively and qualitatively. Additionally, a practical application in dark face detection is performed, preliminarily confirming the potential and feasibility of our method in real-world scenarios.

II. Related Work

From the perspective of training strategies, deep-learning-based LLIE methods can be further partitioned into the following two subclasses.

A. Supervised LLIE

Supervised LLIE means a paired training dataset is required during the training phase. Since this supervised paradigm is

relatively simple, it has been adopted in most early attempts. For example, LLNet [15] and MBLLEN [16] were trained in a similar way using purely synthetic low/normal-light image pairs, simulated via gamma correction and Gaussian noise. Recognizing the limitations of synthetic data, Wei et al. [23] presented a seminal work in which they built the first paired dataset (LOL) that was collected in real scenes by changing exposure time and ISO. Besides, they also proposed a renowned Retinex-related network (Retinex-Net) and trained it using the supervision of LOL. Concurrently, Cai et al. [24] and Chen et al. [25] constructed multi-exposure image datasets, SICE and SID, respectively. Both datasets contain hundreds of low-contrast image sequences captured in real scenes with different exposure levels, and each scene is additionally provided with a corresponding high-quality image for reference or evaluation.

Given the real paired datasets, many efforts have been made to design supervised LLIE networks. As previously mentioned, most of these networks are Retinex-related. Zhang et al. [26] introduced a Retinex-related network (KinD) for kindling the darkness and further improved its enhancement quality in [27]. Wang et al. [28] presented a deep underexposed photo enhancer (DeepUPE) with a strong emphasis on the mapping of illumination. Wu et al. [29] proposed a Retinex-related deep unfolding network (URetinex-Net) that unfolds the optimization problem into a learnable network. More recently, on account of the unsatisfactory quantity of existing datasets (*e.g.*, LOL only has 500 pairs), Hai et al. [30] created a large-scale real-world dataset (LSRW) containing 5,650 pairs of low/normal-light images. Based on the training on LSRW, they also suggested a Retinex-related real-low to real-normal network (R2RNet) to improve image contrast while preserving more details.

In short, despite its simplicity, supervised LLIE requires a paired dataset during the training phase, and the performance of these models strongly relies on the quality and quantity of the dataset used.

B. Unsupervised LLIE

Although several paired training datasets have been constructed, creating these datasets is commonly viewed as an extremely challenging and labor-intensive task. Further realizing the overfitting and model restriction issues caused by paired datasets, serval recent studies have started to tackle the LLIE problem without paired data supervision. As an innovative endeavor, Jiang et al. [31] proposed an effective unsupervised framework based on generative adversarial networks (GANs) [32], [33], [34], and named it EnlightenGAN. The main advantage of this approach is its ability to be trained on unpaired low/normal-light images (*i.e.*, unpaired datasets) with the assistance of a global-local discriminator structure and a self-regularized loss function. Later, Ni et al. [35] presented an unsupervised image enhancement GAN (UEGAN) for further improving the aesthetic quality of generated images.

Moreover, a handful of very recent works have reconsidered solving the LLIE problem in a more appealing way, namely, zero-reference (or zero-shot) LLIE, which is also the main focus of this paper. Zero reference indicates that neither paired nor unpaired datasets are available during training. Following the taxonomy presented in [13] and [22], we group and review zero-reference and zero-shot LLIE models together, despite their subtle differences actually. As a pioneering effort, Zhang et al. [36] introduced a small image-specific exposure correction network (ExCNet) for zero-shot restoration of backlit images. A compelling feature of ExCNet is that it requires neither prior image examples nor prior training. Similarly, motivated by the robust Retinex model [9], Zhu et al. [37] proposed another zero-shot model, namely robust Retinex decomposition network (RRDNet), for underexposed image restoration. Unlike the original Retinex model which involves only illumination and reflectance, RRDNet also takes a noise component into consideration when decomposing the input, and accordingly, it can prevent the noise from being amplified during image contrast stretching.

Meanwhile, a novel zero-reference deep curve estimation (Zero-DCE) method was proposed in [38]. Its basic concept is to transform the light enhancement task from image-to-image mapping to image-specific curve estimation. These parametric curves are estimated by a trainable network optimized using a specific set of non-reference loss functions. With the estimated curves, the final enhanced result is achieved by making pixel-wise adjustments to the dynamic range. An accelerated and lightweight version was later upgraded in [39]. In addition, Liu et al. [40] developed a Retinex-inspired unrolling with architecture search (RUAS) framework to construct and update their enhancement network. Ma et al. [41] designed a cascaded illumination learning process for achieving fast, flexible, and robust implementation. While most previous LLIE models take a single low-light image as input, Fu et al. [42] suggested using paired low-light images. Accordingly, they designed a Retinex-related structure named PairLIE and trained it using carefully selected low-light pairs. Although PairLIE can learn adaptive constraints from both low-light inputs and provide promising results, collecting such paired training datasets is quite laborious and expensive compared to capturing a single low-light image, thereby hindering its wider application.

In summary, unsupervised LLIE eliminates dependence on paired datasets by utilizing unpaired training. In comparison, zero-reference LLIE is particularly appealing and promising as it further mitigates the risk of overfitting and generalizes effectively across various lighting conditions. More importantly, in addition to the learning framework analyzed in the previous section, we believe that the following two points are crucial to zero-reference LLIE and should be appropriately considered in the design of our method: 1) a detailed network structure with thoughtful design, and 2) a well-matched set of non-reference loss functions that can indirectly assess enhancement quality.

III. Proposed Method

In this section, we first introduce the concept of residual quotient learning for LLIE, and then detail the specific structure and zero-reference loss function employed.

A. Residual Quotient Learning for LLIE

According to Retinex theory [11], a low-light image $\boldsymbol{y}$ can be decomposed into two components. That is,

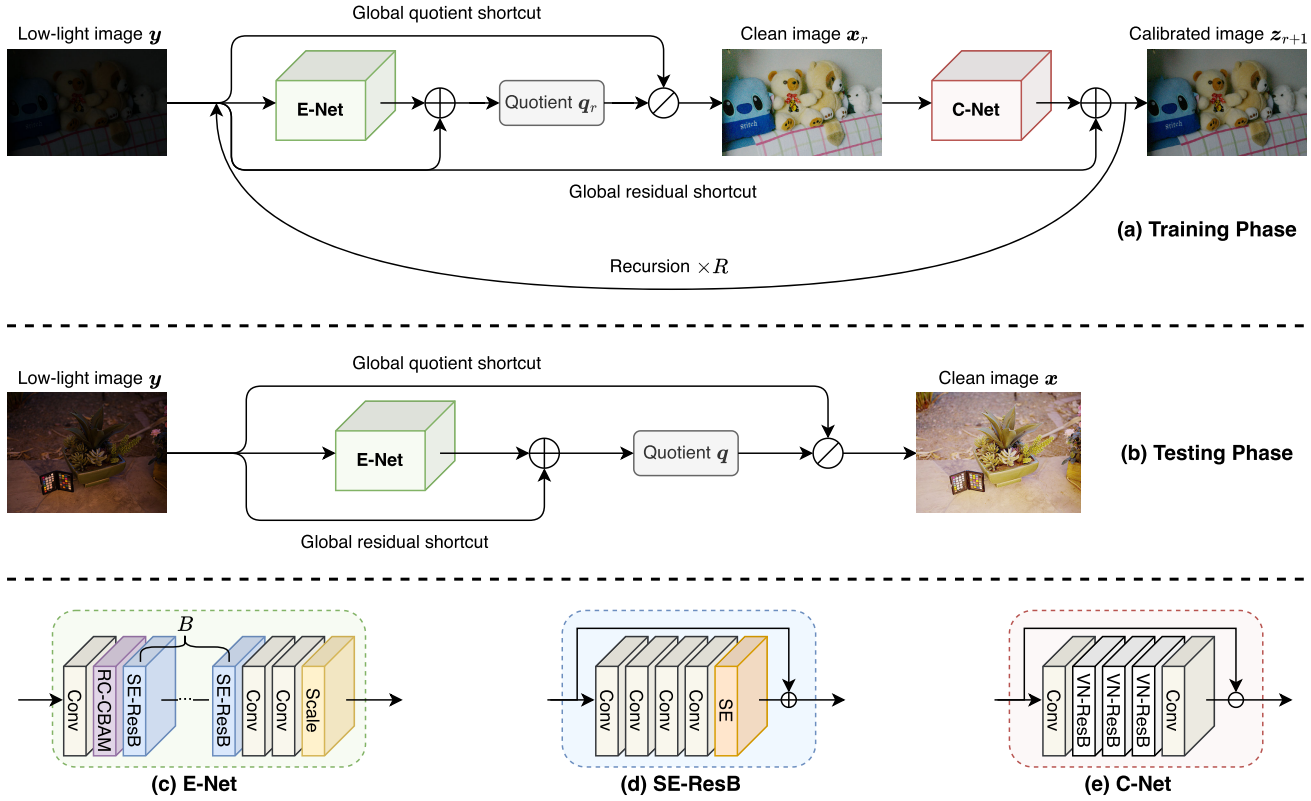


Fig. 2. The overall architecture of the proposed ResQ-Net. Please note that the normalization and activation after each "Conv" layer are omitted for the sake of clarity. The notations $\oplus$, $\oslash$, and $\ominus$ denote the element-wise summation, division, and subtraction, respectively. Zoom in for the best view.

$\boldsymbol{y} = \boldsymbol{x} \otimes \boldsymbol{q}$, where $\otimes$ denotes element-wise multiplication, $\boldsymbol{q}$ represents illumination, indicating the scene lighting intensity, and $\boldsymbol{x}$ is reflectance, describing the intrinsic properties of objects. Recall that most existing Retinex-related methods, such as [23], [26], [29], [37], and [42], tend to enhance the decomposed components individually and then merge them back together to obtain the final enhanced result.

However, we argue that this common practice is not ideal, especially when dealing with natural low-light images with non-uniform degradation. Motivated by [7], [17], and [40], in this paper, we directly treat $\boldsymbol{y}$ and $\boldsymbol{x}$ as the degraded observation and the desired recovery, respectively. That is to say, they are the input and the output of our whole system. The key connection between them is the illumination which we recast as the quotient $\boldsymbol{q}$ of $\boldsymbol{y}$ divided by $\boldsymbol{x}$. If the unknown quotient $\boldsymbol{q}$ can be predicted accurately and precisely, the enhancement problem will be effectively addressed. To this end, we will pay more attention to the modeling of $\boldsymbol{q}$ in our subsequent design.

Actually, such kind of exploration has already been conducted and included in the aforesaid Retinex-related models [23], [26], [29], [37], [42], as one of their main procedures. However, following the widely accepted practice in [5] and [7], all of them implicitly assume that the three color channels of the input image $\boldsymbol{y}$ share the same light intensity map $\boldsymbol{q}$. Put differently, they shrink the channel number of $\boldsymbol{q}$ from three (*i.e.*, color) to one (*i.e.*, grayscale) for model simplicity so that the initial value of $\boldsymbol{q}$ can be estimated by finding the pixel-wise maximum value from the R, G, and B channels of $\boldsymbol{y}$. Nevertheless, we believe that this channel shrinkage operation imposes severe restrictions on the representational capacity of $\boldsymbol{q}$, thereby limiting the enhancement performance, particularly for complex non-uniform illumination.

On the contrary, inspired by the critical discovery [38] that three-channel adjustment can better preserve the inherent color and reduce the risk of over-saturation, we decide to restore the three-channel structure of the quotient $\boldsymbol{q}$. Next, to learn the underlying mapping to $\boldsymbol{q}$, we seek the help of neural networks, and choose to initially take the low-light image $\boldsymbol{y}$ as input for the reason that we intuitively think $\boldsymbol{y}$ and $\boldsymbol{q}$ are visually quite analogous to each other. So we hypothesize that there exists a simpler connection between them and this connection is much easier to optimize. Besides, driven by the tremendous success of residual learning [19], we reformulate the direct mapping to the quotient $\boldsymbol{q}$ as learning its residual function with reference to the model input $\boldsymbol{y}$. In this way, it will further facilitate the subsequent network training and even improve the overall model capacity.

In a nutshell, our overall residual quotient learning framework mentioned above is illustrated in Fig. 1(c), and it can be logically formulated as

$$\text{SystemOutput} = \boldsymbol{x} = \boldsymbol{y} \oslash \boldsymbol{q} := \boldsymbol{y} \oslash (\text{Net}(\boldsymbol{y}) \oplus \boldsymbol{y}) \quad (1)$$

where $\oslash$ and $\oplus$ denote element-wise division and element-wise summation induced by the global quotient shortcut and the global residual shortcut, respectively, and $\text{Net}(\cdot)$ is the only internal network that requires further design and optimization, which will be elaborated on in the next part.

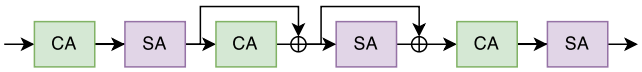


Fig. 3. The detailed structure of RC-CBAM in Fig. 2. Please note that "CA" and "SA" represent the channel attention and spatial attention modules in [45].

B. ResQ-Net

When designing our network, we take both efficiency and effectiveness into consideration. Inspired by the stage-wise network design in [40], [43], and [44], we present a progressive and recursive overall architecture, as illustrated in Fig. 2. One can see that our network is primarily composed of two functional subnets, *i.e.*, the enhancement net (*E-Net*) and the calibration net (*C-Net*), and the paired subnets are stacked R times recursively to form the final structure used in the training phase. Each of them is described in detail below.

1) E-Net: Recall that the role of this part is to learn the underlying mapping from the input low-light image $\boldsymbol{y}$ to the latent quotient $\boldsymbol{q}$ using a residual learning manner. Suppose that the E-Net is in the *r-th* recursion and its current input is denoted as $\boldsymbol{z}_r$. Accordingly, this enhancement process can be expressed as

$$\boldsymbol{q}_r = \text{E-Net}(\boldsymbol{z}_r) \oplus \boldsymbol{y}, \quad \text{s.t.,}\ \boldsymbol{z}_1 = \boldsymbol{y} \tag{2}$$

where $\boldsymbol{z}_1$ means the initial input of E-Net.

Next, considering that natural low-light images are often taken under spatially varying lighting conditions, particularly in real-world scenes, the resulting low-light observations are typically unevenly degraded (see Fig. 4(a) for example). For this reason, they should undergo non-uniform enhancement to yield superior results. Therefore, when designing the detailed structure of E-Net, we take this critical point into account, and accordingly, incorporate the attention mechanisms into our E-Net design. The detailed structure is shown in Fig. 2(c).

Specifically, this subnet consists of three parts: 1) *The head* is constructed for shallow feature extraction, which contains a Conv (short for the sequential arrangement of Conv2d, BN, and ReLU hereinafter) followed by an RC-CBAM module. Our RC-CBAM is a variant of the original CBAM [45] in that it is residually cascaded (see Fig. 3 for details). 2) *The body* is designed for deep feature embedding, which has stacked B SE residual blocks (SE-ResB) [46]. The detailed schema of SE-ResB is depicted in 2(d). 3) *The tail* is established for image reconstruction, which first applies 2 Convs to transform the features back into an image and then globally scales it using a trainable parameter to obtain the predicted residual finally. Note that the remaining configuration of E-Net will be fully investigated in Section IV-B.

2) C-Net: To facilitate and accelerate the subsequent training process, we further adopt a calibration net (C-Net), following the proven practice in [41]. The C-Net is then integrated into the aforementioned residual quotient learning framework, as shown in Fig. 2(a), to record variations in the model output and produce a calibrated image. Similarly, assume that the C-Net is also in the *r-th* recursion and its current input is actually the *r-th* clean image, denoted as $\boldsymbol{x}_r$, and accordingly, this calibration process can be formally outlined as

$$\boldsymbol{z}_{r+1} = \text{C-Net}(\boldsymbol{x}_r) \oplus \boldsymbol{y} \tag{3}$$

where $\boldsymbol{z}_{r+1}$ is the $(r{+}1)$-*th* calibrated image, and note that it will serve as the new input in the next recursion.

The internal structure of our C-Net is provided in Fig. 2(e). Akin to E-Net, it consists of three parts (i.e., the *head*, *body*, and *tail*) and an attached shortcut for element-wise subtraction. But, in contrast, the three parts comprise a Conv, 3 stacked vanilla residual blocks (VN-ResB) [19], and another Conv, respectively. The benefits of introducing C-Net are threefold: Firstly, it significantly expedites the entire training process. Secondly, it improves the robustness of our ResQ-Net to more various degradation levels, thus optimizing the overall performance. Lastly, as training proceeds, it ensures that the outputs of E-Net at each recursion converge to almost the same value, thereby stabilizing the whole system.

In summary, by virtue of the above elaborately designed architecture, the proposed ResQ-Net can be efficiently trained using the progressive and recursive structure, as illustrated in Fig. 2(a). Meanwhile, because of the convergence across all recursions, this structure can be further streamlined during the testing phase to accelerate the inference time, just as depicted in Fig. 2(b).

C. Zero-Reference Loss Function

To fully enable zero-reference training, our ResQ-Net needs to be further optimized based on the criteria established by a series of non-reference losses. In this part, we propose a carefully designed zero-reference loss function, which consists of the following four terms.

1) Fidelity Loss: This criterion is adopted to measure the pixel-level consistency between the input of E-Net and the estimated quotient at each recursion. Accordingly, it can be formulated as

$$\mathcal{L}_F = \sum_{r=1}^{R} \left\| \boldsymbol{z}_r - \boldsymbol{q}_r \right\|_2^2 \tag{4}$$

where $\boldsymbol{z}_r$ and $\boldsymbol{q}_r$ denote the input of E-Net and the estimated quotient in the *r-th* recursion, respectively, and R is the total number of recursions.

2) Smoothness Loss: Recall that the quotient $\boldsymbol{q}$ physically represents the scene light intensity map, so it should be piece-wise smooth and textureless. To eliminate undesired image patterns, a smoothness loss needs to be introduced to penalize the color differences between adjacent pixels. Inspired by the spatially-variant flattening criterion [47], this term is formally defined as

$$\mathcal{L}_S = \sum_{r=1}^{R} \sum_{n=1}^{N} \sum_{i \in \mathcal{N}_5(n)} w_{r,(n,i)} \left\| \boldsymbol{q}_{r,n} - \boldsymbol{q}_{r,i} \right\|_1 \tag{5}$$

where $\boldsymbol{q}_{r,n}$ denotes *n-th* pixel value in $\boldsymbol{q}_r$, N is the total pixel number, $\mathcal{N}_5(n)$ represents the neighbouring pixels of n in its 5×5 window, and $w_{r,(n,i)}$ is the corresponding weight that is calculated as

$$w_{r,(n,i)} = \exp\left(-\frac{\left\| \boldsymbol{z}_{r,n} - \boldsymbol{z}_{r,i} \right\|_2^2}{2\sigma^2} \right) \tag{6}$$

where σ is the standard deviation for the Gaussian kernel and is set to 0.1. Note that the pixel values $z_{r,n}$ and $z_{r,i}$ are in the YCbCr color space.

3) Color Loss: Based on the Gray-World color constancy hypothesis [48], which states that the average values of the three channels in a natural color image statistically approximate the same gray level, we deploy a color constancy loss to correct the possible color casts in our system output at each recursion. Following [39], the term can be written as

$$\mathcal{L}_C = \sum_{r=1}^{R} \sum_{(p,q)\in\varepsilon} (J_r^p - J_r^q)^2, \quad \varepsilon = \{\mathrm{R},\ \mathrm{G},\ \mathrm{B}\} \tag{7}$$

where J_r^p denotes the average intensity value of p channel in the enhanced image in the *r-th* recursion, (p, q) represents a pair of channels.

4) Perceptual Loss: As a necessary complement to pixel-level fidelity losses, perceptual losses [49] have proven effective in measuring perceptual similarity, and have therefore been widely applied to many low-level computer vision tasks [50], [51]. However, the existing form cannot be directly employed here, since no ground truth is available in our zero-reference training setting. To address this, we reformulate it into the following applicable variant

$$\mathcal{L}_P = \sum_{r=1}^{R} \left\| \phi_l(z_r) - \phi_l(q_r) \right\|_2^2 \tag{8}$$

where $\phi_l(\cdot)$ refers to the intermediate feature of the *l-th* layer in the VGG16 [52] network pretrained on ImageNet [53] given a specific input, and we set $l =$ relu4_3 in this paper.

Overall Loss Function: In summary, our overall loss function can be expressed as a linear combination of all losses

$$\mathcal{L}_{All} = \lambda_F \mathcal{L}_F + \lambda_S \mathcal{L}_S + \lambda_C \mathcal{L}_C + \lambda_P \mathcal{L}_P \tag{9}$$

where λ_F, λ_S, λ_C, and λ_P represent the weights corresponding to each loss and are empirically set to 1.5, 1, 0.2, and 1, respectively.

IV. Experimental Results

In this section, we thoroughly explore and verify the performance of our proposed model. First, the implementation details are introduced, particularly the training and testing configurations. Next, we conduct an ablation study to confirm the validity of the proposed framework. Then, we quantitatively and qualitatively evaluate our trained model in comparison with recent state-of-the-art counterparts on benchmark datasets. Finally, we perform a practical application in dark face detection to enhance test diversity and demonstrate the practicality of our method.

A. Implementation Details

In the training phase, we use the union of 500 randomly sampled low-light images from the training part of LSRW [30] and 500 randomly sampled low-light images from the MIT-Adobe FiveK Dataset [54] as our training dataset. The proposed ResQ-Net is implemented with PyTorch, using the Adam [55] optimizer with a learning rate of 0.0001 and other parameters set by default. The batch size is set to 4 and the other main hyper-parameters of our model are empirically set as follows: $R = 3$, $B = 3$, $L = 4$, $C = 3$. We train our final model for 400 epochs and select the point with the best performance.

In the testing phase, we evaluate our trained model on both paired and unpaired benchmark datasets. Specifically, for reference evaluation, the testing part of LSRW (50 pairs) and another 115 randomly sampled image pairs from the MIT-Adobe FiveK Dataset (MIT for short hereinafter) are selected. Additionally, the testing part of LOL v1 [23] and LOL v2 [56] (115 pairs, LOL v1+v2 for short) and LOL Synthetic [56] (100 pairs, LOL SYN for short) are included to further enhance diversity. Meanwhile, for no-reference evaluation, five widely accepted testing datasets are employed, namely, MEF [57], DICM [58], LIME [7], Fusion [23], and VV.[1] As for evaluation metrics, four full-reference indicators (*i.e.*, PSNR, SSIM [59], LOE [5], and LPIPS [60]) and one no-reference indicator (*i.e.*, NIQE [61]) are adopted. Finally, all experiments are conducted on an Ubuntu server equipped with 2 NVIDIA GeForce RTX 4090 GPUs.

TABLE I

Ablation Study on the Architecture of E-Net. The Test Is Conducted on the MIT Dataset. The Best Performance Is in Bold

Architecture of E-Net			Metrics		
SE-ResB	CBAM	RC-CBAM	PSNR↑	SSIM↑	LPIPS↓
×	×	×	19.45	0.8927	0.1306
✓	×	×	20.81	0.9114	0.1139
×	✓	×	18.15	0.8432	0.2183
×	×	✓	20.52	0.9090	0.1138
✓	✓	×	16.99	0.8596	0.1657
✓	×	✓	**20.87**	**0.9152**	**0.1069**

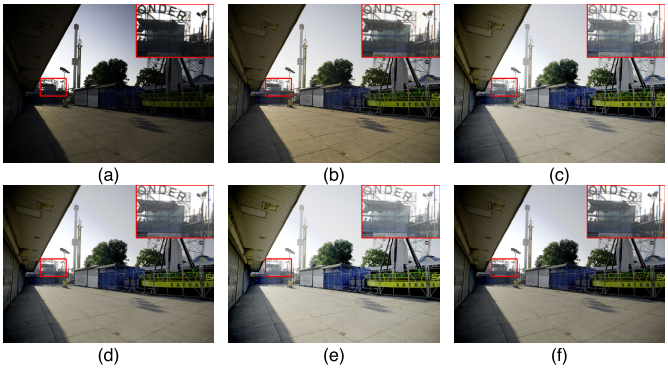


Fig. 4. Visual comparison of different attentions in E-Net. (a) Low-light input. (b) Reference (PSNR/SSIM). (c) w/o any attention (19.08/0.8963). (d) w/ SE (26.06/0.9414). (e) w/ SE & CABM (15.46/0.8361). (f) w/ SE & RC-CABM (27.81/0.9457). Zoom in to see the differences in the highlighted regions. Best viewed in color.

B. Ablation Study

To validate the effectiveness of the proposed modules, we conduct a series of ablation studies on our model. For the

[1] https://sites.google.com/site/vonikakis/datasets

TABLE II

ABLATION STUDY ON THE CONFIGURATION OF SE-RESB. THE TEST IS CONDUCTED ON THE MIT DATASET. THE BEST PERFORMANCE IS IN BOLD

Configuration of SE-ResB			Metrics		
# Blocks	# Layers	# Channels	PSNR↑	SSIM↑	LPIPS↓
3	2	3	20.03	0.8925	0.1305
3	**4**	3	**20.87**	**0.9152**	**0.1069**
3	6	3	19.42	0.8872	0.1331
2	4	3	20.73	0.9039	0.1191
3	4	3	**20.87**	**0.9152**	**0.1069**
4	4	3	20.61	0.8896	0.1368
3	4	**3**	**20.87**	**0.9152**	**0.1069**
3	4	8	20.71	0.9006	0.1228

TABLE III

ABLATION STUDY ON THE CONTRIBUTION OF EACH LOSS. THE TEST IS CONDUCTED ON THE MIT DATASET. THE BEST PERFORMANCE IS IN BOLD

Contribution of Each Loss				Metrics		
$\mathcal{L}_F$	$\mathcal{L}_S$	$\mathcal{L}_C$	$\mathcal{L}_P$	PSNR↑	SSIM↑	LPIPS↓
✓	×	×	×	15.18	0.8339	0.1959
✓	✓	×	×	19.23	0.8848	0.1492
✓	✓	✓	×	20.64	0.9002	0.1328
✓	✓	✓	✓	**20.87**	**0.9152**	**0.1069**

sake of simplicity, the following tests are performed exclusively on the MIT dataset, and the respective performances are quantitatively evaluated using the combination of PSNR, SSIM, and LPIPS.

1) Effectiveness of the Attention Mechanisms in E-Net: We begin by analyzing the impact of the attention mechanisms introduced in E-Net. The quantitative comparison is reported in Table I. Specifically, the widely used ResNet structure (*i.e.*, the plain residual blocks without any attention) is preliminarily utilized as the baseline. Given the benefits of SE attention, we first replace the plain ResBlocks with the advanced SE-ResBs, as depicted in Fig. 2(d). It can be seen that this change results in a significant improvement in overall performance. However, when we continue adding the CBAM attention to the head part of E-Net, the performance drops dramatically. This phenomenon suggests a serious conflict between the original CBAM and SE attention. To address this, we modify the original CBAM structure by cascading the CA and SA modules in a residual manner, as illustrated in Fig. 3. Fortunately, the resulting RC-CBAM collaborates effectively with the SE-ResB and gives the best overall performance.

In addition, a visual comparison is illustrated in Fig. 4. It can be noted that the low-light input is captured under real-world conditions and its content is typically non-uniformly degraded (*e.g.*, the sky versus the shadows). According to the reference image retouched by a photography expert, we can observe that the model with full attention mechanisms produces the most visually pleasing output. It avoids overexposure, minimizes color distortion, and maintains sharp edges, outperforming the other variants in these aspects. Therefore, we can conclude that the visual observations are consistent with the previous

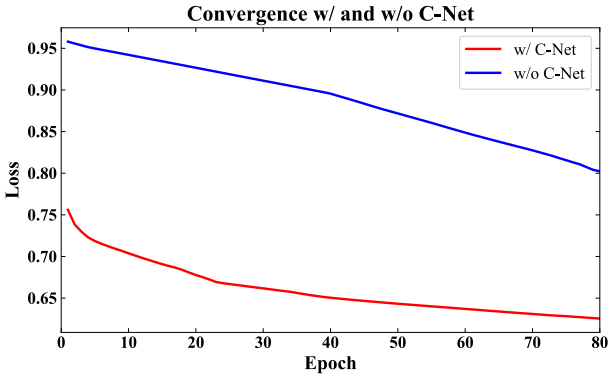


Fig. 5. Impact of C-Net on model training. Note that a reduced epoch number is used here to save time.

quantitative comparisons, confirming the effectiveness of the attention mechanisms introduced in this paper.

2) Configuration of SE-ResB: We then conduct experiments to determine the detailed configuration of SE-ResB, including the number of SE-ResBs, the number of layers within each block, and the number of input and output channels. These hyper-parameters are denoted by B, L, and C, respectively. Based on our preliminary study, the entire test is divided into three groups, and all corresponding results are listed in Table II. Specifically, in the first group, we keep B and C constant, while varying L across several values. It can be seen that our model achieves the best performance when $L = 4$. Similarly, we employ the same approach to determine the optimal values of the other two parameters in the second and third groups of experiments, respectively. In conclusion, we use $B = 3$, $L = 4$, and $C = 3$ as our default SE-ResB configuration, since this setup provides the best performance.

3) Effectiveness of C-Net: Next, to verify the effectiveness of C-Net, we assess the impact of removing it from the full model on subsequent training. To this end, the model without C-Net is additionally trained using identical settings. The training procedures are recorded and compared with those of the full model. The details are plotted in Fig. 5, from which we can see that the model with C-Net consistently achieves a faster convergence rate and a lower training error, demonstrating the effectiveness and necessity of introducing C-Net into our proposed framework.

4) Contribution of Each Loss: Finally, we evaluate the contribution of each individual loss term. To this end, we additionally train three variant models, each omitting one or more loss terms during training, and compare them with the full model whose loss function includes all four terms, *i.e.*, $\mathcal{L}_F$, $\mathcal{L}_S$, $\mathcal{L}_C$, and $\mathcal{L}_P$. To ensure consistency during training and testing, we maintain identical settings across all models. The corresponding results are presented in Table III. It can be seen that our model's performance improves progressively as more loss terms are introduced, with the full model yielding the best results. This confirms both the effectiveness and necessity of each loss term introduced in our system.

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

In this subsection, we quantitatively and qualitatively compare our proposed ResQ-Net with 17 state-of-the-art

TABLE IV

QUANTITATIVE EVALUATION OF COMPARED METHODS ON TWO REFERENCE BENCHMARK DATASETS. THE BEST PERFORMANCE IS IN BOLD, AND THE SECOND-BEST PERFORMANCE IS UNDERLINED. "C", "S", AND "U" REPRESENT CONVENTIONAL, SUPERVISED, AND UNSUPERVISED METHODS, RESPECTIVELY

Method	Type	LSRW					MIT				
		PSNR↑	SSIM↑	LOE↓	LPIPS↓	NIQE↓	PSNR↑	SSIM↑	LOE↓	LPIPS↓	NIQE↓
HE	C	13.06	0.3097	322.8	0.4221	3.938	15.93	0.7033	752.3	0.4519	5.938
BIMEF [8]		15.03	0.4754	251.8	0.3268	3.834	11.72	0.4607	1280.2	0.1601	4.742
LIME [7]		15.68	0.3766	369.0	0.3538	3.873	10.18	0.3924	1350.8	0.1881	4.522
NPE [5]		16.21	0.3989	542.0	0.3686	3.761	11.64	0.4386	1306.3	0.1739	4.498
SRIE [6]		13.35	0.4225	346.0	0.3400	3.816	11.41	0.4382	1291.6	0.1601	4.608
LECARM [10]		15.34	0.4262	306.0	0.3268	10.511	17.47	0.8338	636.2	0.2273	4.362
Retinex-Net [23]	S	15.49	0.3546	629.6	0.4322	4.146	14.73	0.7377	1629.6	0.3816	4.750
MBLLEN [16]		16.52	0.4867	257.1	0.3834	4.722	18.01	0.7506	1257.1	0.2830	4.169
KinD [26]		17.15	0.5200	422.8	0.4186	3.851	17.84	0.7683	1312.7	0.2503	4.378
URetinex-Net [29]		18.27	**0.5259**	280.8	0.3190	4.180	18.56	0.8222	680.8	0.1881	4.152
RRDNet [37]	U	12.66	0.3862	259.7	0.3917	4.353	19.60	0.8302	631.5	0.2291	4.397
EnlightenGAN [31]		17.59	0.4794	431.4	0.3122	3.995	15.56	0.8002	789.4	0.2127	4.233
UEGAN [35]		9.97	0.2047	354.4	0.4830	4.802	19.54	0.8698	267.1	0.1988	4.287
ZeroDCE [38]		16.26	0.4634	302.1	0.3278	3.764	17.96	0.8429	602.1	0.2257	4.244
RUAS [40]		14.03	0.4028	346.1	0.3852	4.240	8.46	0.5403	1346.1	0.5850	9.233
SCI [41]		15.24	0.4240	273.8	0.3221	3.926	19.52	0.8845	374.0	0.1762	4.167
PairLIE [42]		17.60	0.5117	309.0	0.3288	3.698	14.47	0.7594	383.9	0.1976	**3.933**
ResQ-Net (Ours)		**18.63**	0.4709	**251.7**	**0.3083**	**3.686**	**20.87**	**0.9152**	**196.1**	**0.1069**	4.234

TABLE V

QUANTITATIVE EVALUATION OF COMPARED METHODS ON THE OTHER TWO REFERENCE BENCHMARK DATASETS. THE BEST PERFORMANCE IS IN BOLD, AND THE SECOND-BEST PERFORMANCE IS UNDERLINED. "C", "S", AND "U" REPRESENT CONVENTIONAL, SUPERVISED, AND UNSUPERVISED METHODS, RESPECTIVELY

Method	Type	LOL v1+v2					LOL SYN				
		PSNR↑	SSIM↑	LOE↓	LPIPS↓	NIQE↓	PSNR↑	SSIM↑	LOE↓	LPIPS↓	NIQE↓
HE	C	11.88	0.3437	320.4	0.6027	9.305	12.89	0.2785	273.3	0.8389	12.314
BIMEF [8]		17.19	0.6749	239.3	0.2799	4.517	12.76	0.4643	253.7	0.6167	6.883
LIME [7]		15.61	0.4759	408.7	0.4430	5.739	15.19	0.3638	389.3	0.7568	9.336
NPE [5]		17.55	0.5250	534.9	0.4486	5.096	15.18	0.3903	394.0	0.7485	8.709
SRIE [6]		13.99	0.5599	362.0	0.3070	**3.727**	11.02	0.3860	308.8	0.6283	6.828
LECARM [10]		16.94	0.5419	204.3	0.3225	8.018	12.49	0.3396	241.7	0.6863	10.597
Retinex-Net [23]	S	15.95	0.3910	566.8	0.5535	9.356	14.04	0.2729	486.5	0.8183	11.568
MBLLEN [16]		17.87	0.6855	166.9	0.2662	4.419	15.78	0.5014	209.9	0.5990	6.531
KinD [26]		18.63	0.7213	389.8	0.2975	4.744	16.31	0.5683	401.8	0.5524	**5.193**
URetinex-Net [29]		**20.60**	**0.8368**	205.4	**0.1277**	4.396	**17.70**	0.6293	221.6	**0.4478**	6.848
RRDNet [37]	U	13.24	0.4944	213.3	0.3152	3.799	10.50	0.3102	238.7	0.7290	7.721
EnlightenGAN [31]		18.54	0.6718	422.6	0.3032	4.836	15.18	0.4819	424.3	0.6197	7.455
UEGAN [35]		12.22	0.2620	206.1	0.4461	5.627	8.90	0.1933	210.1	0.6998	7.239
ZeroDCE [38]		18.06	0.5779	232.1	0.3093	7.997	12.99	0.3645	301.0	0.6839	10.666
RUAS [40]		15.05	0.4562	167.6	0.3716	8.482	12.19	0.2971	259.6	0.7481	11.227
SCI [41]		16.97	0.5320	151.9	0.3120	8.022	12.36	0.3233	220.1	0.7008	10.684
PairLIE [42]		18.09	0.7270	281.7	0.2630	4.451	16.11	0.5661	306.0	0.5644	5.430
ResQ-Net (Ours)		19.71	0.7342	**148.0**	0.2615	3.817	16.94	**0.6389**	**208.7**	0.5503	5.269

LLIE methods, including histogram equalization (HE), BIMEF [8], LIME [7], NPE [5], SRIE [6], LECARM [10], Retinex-Net [23], MBLLEN [16], KinD [26], URetinex-Net [29], RRDNet [37], EnlightenGAN [31], UEGAN [35], ZeroDCE [38], RUAS [40], SCI [41], PairLIE [42]. As introduced before, the first six methods are convectional, the next four methods are supervised, and the remaining methods are unsupervised together with ours. All the source codes are obtained from the respective authors' official repositories, and we directly use the default settings recommended by the authors to ensure optimal performance, unless otherwise stated.

1) Quantitative Comparison: The average performance of the compared methods on the reference datasets (LSRW, MIT, LOL v1+v2, and LOL SYN) is quantitatively summarized in Table IV and Table V. As one can see, by virtue of the advanced network implementations, the deep-learning-based methods are able to achieve much better scores compared with the conventional ones. Meanwhile, it is evident from the comparison that our ResQ-Net consistently provides the

TABLE VI

QUANTITATIVE EVALUATION OF COMPARED METHODS ON FIVE NO-REFERENCE BENCHMARK DATASETS USING NIQE. THE BEST PERFORMANCE IS IN BOLD, AND THE SECOND-BEST PERFORMANCE IS UNDERLINED. "C", "S", AND "U" REPRESENT CONVENTIONAL, SUPERVISED, AND UNSUPERVISED METHODS, RESPECTIVELY

Method	Type	No-Reference Benchmark Dataset					Average
		MEF	DICM	LIME	Fusion	VV	Value
HE	C	4.472	<u>3.407</u>	4.079	3.668	2.984	3.722
BIMEF [8]		3.675	3.910	4.794	3.857	3.123	3.872
LIME [7]		3.758	4.001	<u>4.036</u>	3.663	<u>2.748</u>	3.641
NPE [5]		3.946	3.845	4.879	3.943	3.029	3.928
SRIE [6]		3.680	3.983	4.827	3.692	3.136	3.864
LECARM [10]		3.682	4.040	4.113	3.688	2.921	3.689
Retinex-Net [23]	S	4.395	4.523	4.591	4.158	2.767	4.087
MBLLEN [16]		4.740	3.722	4.627	4.578	3.849	4.303
KinD [26]		4.786	4.132	4.745	5.005	4.237	4.581
URetinex-Net [29]		3.789	3.459	4.341	3.818	3.019	3.685
RRDNet [37]	U	3.781	6.727	6.125	5.781	2.979	5.078
EnlightenGAN [31]		**3.420**	3.568	4.061	<u>3.654</u>	2.823	<u>3.505</u>
UEGAN [35]		5.132	4.046	4.540	4.228	3.696	4.328
ZeroDCE [38]		3.582	3.620	4.764	3.850	3.087	3.780
RUAS [40]		5.109	5.727	4.697	6.080	5.346	5.392
SCI [41]		3.631	5.433	4.180	3.907	2.833	3.997
PairLIE [42]		4.164	3.519	4.515	5.002	3.654	4.171
ResQ-Net (Ours)		<u>3.477</u>	**3.388**	**4.034**	**3.632**	**2.732**	**3.453**

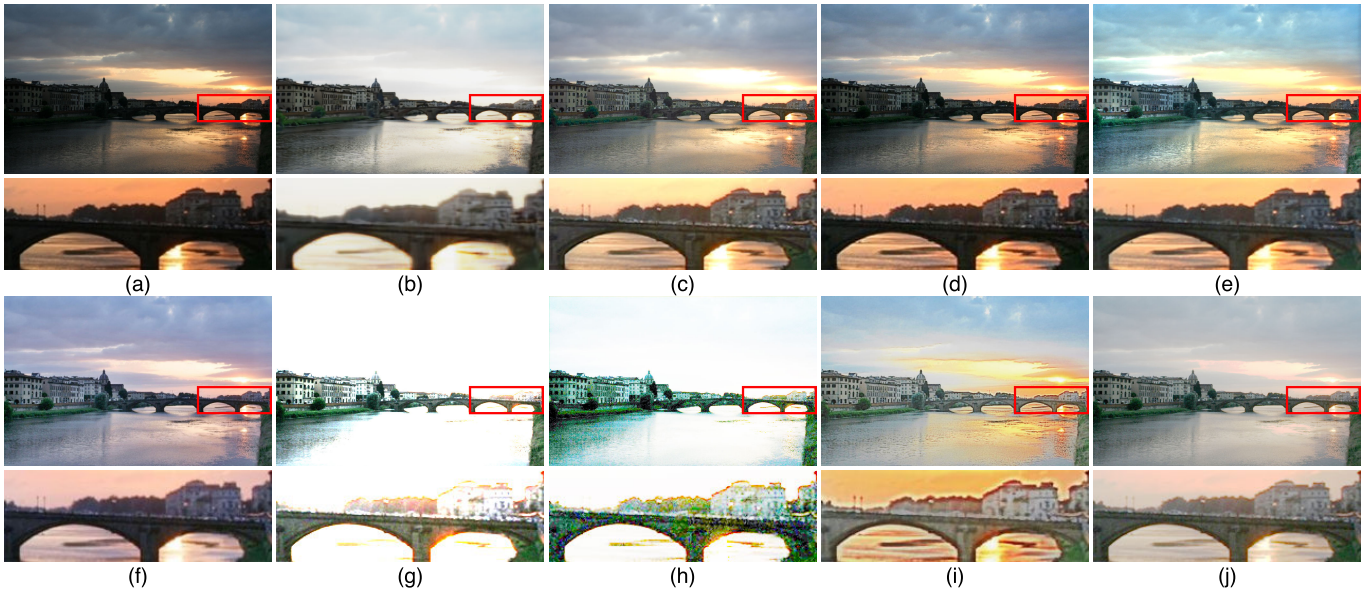


Fig. 6. Visual comparison of the compared methods on a natural non-uniform illumination image from the DICM dataset. (a) Low-light input. (b) KinD [26]. (c) URetinex-Net [29]. (d) RRDNet [37]. (e) EnlightenGAN [31]. (f) ZeroDCE [38]. (g) RUAS [40]. (h) SCI [41]. (i) PairLIE [42]. (j) ResQ-Net (Ours). Note the differences in the enlarged areas. Zoom in for the best view.

best or comparable performance among all the compared methods, mainly due to the positive effects brought about by the proposed residual quotient learning framework and the well-designed network structure.

In addition to the above reference datasets, we further compare these methods on five no-reference benchmark datasets. It is worth noting that only the no-reference metric NIQE is applicable here, as no reference images are provided in these datasets. Table VI reports the detailed numerical results. As we can see, our ResQ-Net continues to perform the best in this experiment, demonstrating its superiority over the compared methods once again.

2) Visual Comparison: Apart from the quantitative evaluation conducted previously, two groups of typical visual comparisons are illustrated in Figs. 6 and 7 to qualitatively compare the enhanced results of the different methods. Note that the methods HE, BIMEF, LIME, NPE, SRIE, LECARM, Retinex-Net, MBLLEN, and UEGAN are excluded from the visual comparisons due to their unsatisfactory quantitative evaluation. It can be observed that our proposed approach produces the most visually pleasing outputs in the sense that it avoids overexposure, minimizes color deviation, and maintains sharp edges and fine details. In contrast, the results enhanced by the other competitors are overexposed and

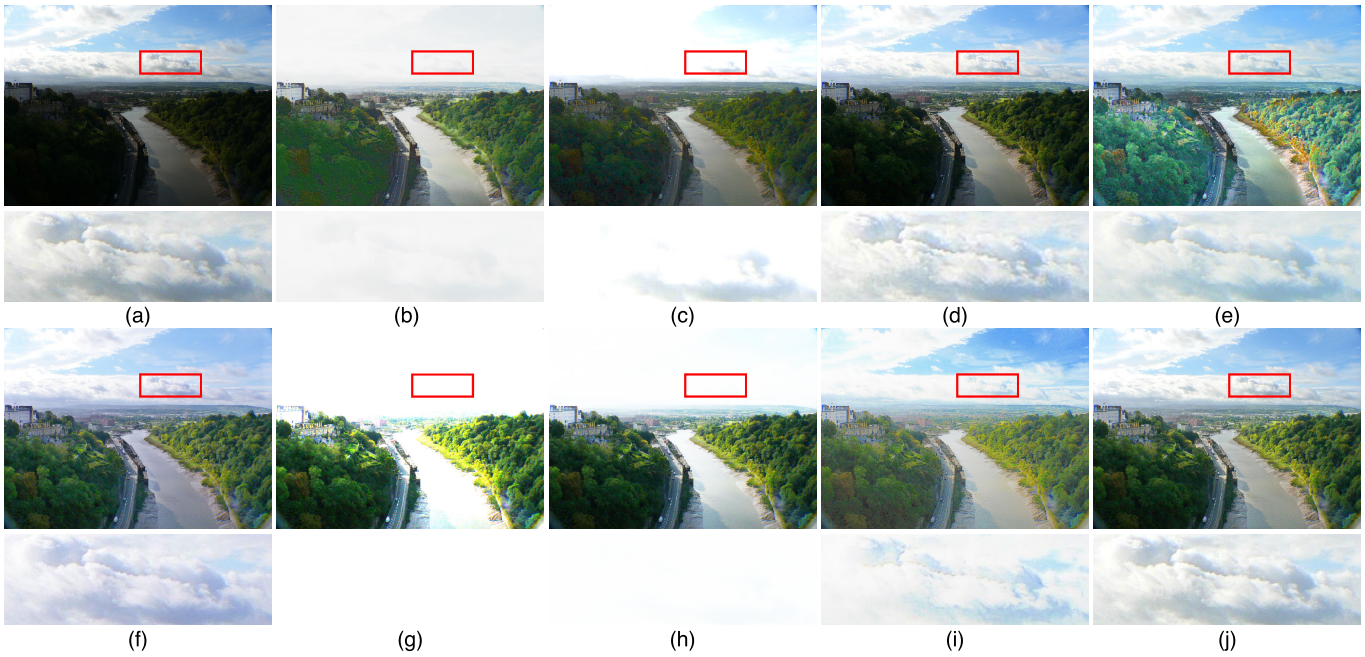


Fig. 7. Visual comparison of the compared methods on a natural non-uniform illumination image from the VV dataset. (a) Low-light input. (b) KinD [26]. (c) URetinex-Net [29]. (d) RRDNet [37]. (e) EnlightenGAN [31]. (f) ZeroDCE [38]. (g) RUAS [40]. (h) SCI [41]. (i) PairLIE [42]. (j) ResQ-Net (Ours). Note that the enlarged area of (a) is within the norm-light region and can therefore be viewed as the label. Zoom in for the best view.

TABLE VII

USER STUDY OF COMPARED METHODS ON VISUAL PERCEPTION.THE SCORING CRITERIA ARE SHARPNESS, BRIGHTNESS, AND COLOR.THE BEST TOPSIS RESULT IS IN BOLD

Method	Type	Scoring Criteria			TOPSIS
		Sharp.	Bright.	Color	Result
HE	C	3.0	2.7	1.2	0.0362
BIMEF [8]		2.8	2.6	2.5	0.0424
LIME [7]		3.9	3.3	3.5	0.0714
NPE [5]		3.4	3.7	3.1	0.0652
SRIE [6]		1.8	2.5	2.3	0.0260
LECARM [10]		3.7	3.4	4.1	0.0734
Retinex-Net [23]	S	1.2	2.3	1.9	0.0146
MBLLEN [16]		3.8	3.6	3.2	0.0706
KinD [26]		2.5	3.6	2.9	0.0496
URetinex-Net [29]		3.3	3.6	3.8	0.0685
RRDNet [37]	U	3.4	2.8	3.0	0.0568
EnlightenGAN [31]		3.6	3.6	3.2	0.0680
UEGAN [35]		3.1	1.8	3.4	0.0497
ZeroDCE [38]		3.7	3.0	3.2	0.0639
RUAS [40]		2.1	5.0	2.3	0.0493
SCI [41]		4.5	3.7	3.3	0.0787
PairLIE [42]		2.2	2.6	2.7	0.0363
ResQ-Net (Ours)		4.6	3.5	3.6	**0.0795**

unnatural to a certain extent, making them far less acceptable than ours.

3) User Study: As an important complement to the visual comparison, a user study is performed to manually quantify the subjective visual quality of the compared methods. Specifically, we randomly select two images from each testing dataset and compare the enhancement effects of various methods on the selected images. Thus, a total of 324 images are collected for this user study. Ten human subjects are then invited to rate these images separately and individually. During the rating process, the enhanced images are displayed on a monitor in random order, and the participants are requested to assign three integer scores ranging from 1 (worst) to 5 (best) to each image based on three scoring criteria: sharpness, brightness, and color. Note that the participants are allowed to zoom in and out for clarity, and no reference images are provided throughout the entire test.

The average subjective scores are listed in Table VII. Based on these scores, we further employ the TOPSIS [62] to determine the final rankings. TOPSIS is a multi-criteria decision-making method which fully utilizes the information from the original data and accurately reflects the differences between the evaluation schemes. When applying the TOPSIS method, we assign equal weight to each scoring factor and normalize the computed values to ensure that their summation equals 1. The final TOPSIS results are presented in the same table. As can be seen, our model achieves the highest TOPSIS result and therefore ranks first. This experiment demonstrates that our enhanced images are most favored by the human testers.

4) Model Complexity: In addition to the performance evaluation conducted above, we make a further comparison on model complexity to assess the efficiency of the compared methods. Note that this assessment focuses exclusively on deep-learning-based methods, since they benefit significantly from GPU acceleration. Table VIII reports the corresponding FLOPs, model parameters, and inference time for each compared methods on the LOL dataset. As we can see, our method ranks third, slightly behind SCI and RUAS, while still positioning itself among the top-performing lightweight LLIE models.

In conclusion, our method strikes a strong balance between model complexity and performance, showcasing its efficiency

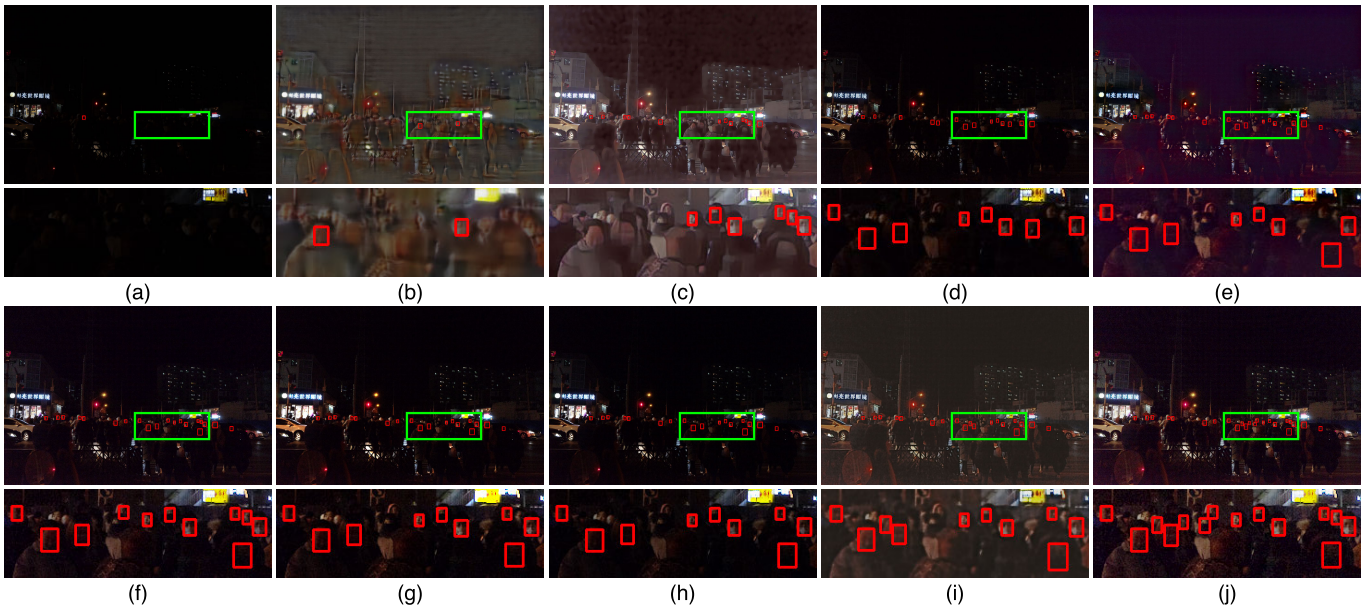


Fig. 8. Dark face detection on a challenging example from the DARK FACE dataset. (a) Raw detection on the low-light input. (b)-(j) Enhanced detection on the results of different LLIE methods: (b) KinD [26]. (c) URetinex-Net [29]. (d) RRDNet [37]. (e) EnlightenGAN [31]. (f) ZeroDCE [38]. (g) RUAS [40]. (h) SCI [41]. (i) PairLIE [42]. (j) ResQ-Net (Ours). Note that for each cell in the figure, the upper image shows the complete detection result with the bounding boxes in red, while the lower image provides an enlarged view of the content within the green box. Zoom in for the best view.

TABLE VIII

QUANTITATIVE EVALUATION OF DEEP-LEARNING-BASED METHODS ON MODEL COMPLEXITY IN TERMS OF FLOPS, PARAMETERS, AND INFERENCE TIME (GPU SECONDS)

Method	Type	Model Complexity		
		FLOPs (G)	Params (M)	Time (S)
Retinex-Net [23]	S	136.02	0.8383	0.0864
MBLLEN [16]		60.02	0.4502	0.2111
KinD [26]		29.13	8.5402	0.1529
URetinex-Net [29]		58.27	0.3621	0.0176
RRDNet [37]	U	30.66	0.1282	1.5677
EnlightenGAN [31]		61.01	8.6360	0.0701
UEGAN [35]		32.72	16.6149	0.0435
ZeroDCE [38]		5.21	0.0789	0.0204
RUAS [40]		0.21	0.0034	0.0165
SCI [41]		0.08	0.0004	0.0160
PairLIE [42]		22.35	0.3418	0.1985
ResQ-Net (Ours)		0.49	0.0123	0.0178

as a lightweight LLIE architecture while consistently delivering superior low-light enhancement performance.

D. Real Application in Dark Face Detection

To finally evaluate the effectiveness and practicality of the compared methods, we conduct another set of experiments on a typical application of LLIE, namely, face detection in dark environments. As we all know, face detection is a fundamental task in high-level computer vision. However, this task becomes extremely challenging under low-light conditions due to poor visibility.

The subsequent procedures mainly follow the widely used practice in [22]. Specifically, the experiments are conducted on the DARK FACE [63] dataset, which provides real-world low-light images captured during the nighttime. Since the testing set is publicly unavailable, the evaluation is performed on 500 images randomly sampled from the training and validation portions. As recommended, the dual shot face detector (DSFD) [64], pretrained on the WIDER FACE [65] dataset, is used as the standard face detector. Both low-light images and their corresponding enhanced results of different LLIE methods are fed into DSFD for raw and enhanced face detection. Then, we computer and compare the average precision (AP) under different IoU thresholds (0.5, 0.6, and 0.7) using the official evaluation tool[2] as well as the accuracy rate (ACC).

TABLE IX

QUANTITATIVE EVALUATION OF COMPARED METHODS ON THE DARKFACE DATASET. THE BEST PERFORMANCE IS IN BOLD

Method	Type	AP↑ Under Different IoUs			ACC↑
		0.5	0.6	0.7	
Low-Light Input	-	0.1237	0.0721	0.0046	0.1462
HE	C	0.1426	0.0998	0.0137	0.2151
BIMEF [8]		0.2282	0.1226	0.0141	0.3143
LIME [7]		0.1799	0.0902	0.0035	0.2960
NPE [5]		0.1606	0.1085	0.0025	0.2499
SRIE [6]		0.1547	0.0874	0.0101	0.2366
LECARM [10]		0.1996	0.0870	0.0087	0.2772
Retinex-Net [23]	S	0.2492	0.1265	0.0151	0.3146
MBLLEN [16]		0.2125	0.1304	0.0128	0.2474
KinD [26]		0.1089	0.0753	0.0127	0.1512
URetinex-Net [29]		0.2203	0.1118	0.0078	0.2913
RRDNet [37]	U	0.1779	0.0716	0.0169	0.2474
EnlightenGAN [31]		0.2128	0.1033	0.0098	0.2947
UEGAN [35]		0.1020	0.0508	0.0061	0.1495
ZeroDCE [38]		0.2330	0.0981	0.0075	0.3113
RUAS [40]		0.2376	0.1037	0.0111	0.3014
SCI [41]		0.2357	0.0994	0.0084	0.2878
PairLIE [42]		0.2288	0.0971	0.0103	0.3113
ResQ-Net (Ours)		**0.2582**	**0.1330**	**0.0182**	**0.3303**

[2]https://github.com/Ir1d/DARKFACE_eval_tools

Table IX summarizes all the face detection results. We can observe a significant overall increase in detection performance (precision and accuracy) after pre-enhancing the raw low-light images using various LLIE methods. This confirms the benefits of LLIE for dark face detection and highlights its potential applications in this area. Additionally, among all the compared LLIE methods, our proposed model surpasses its competitor in all four detection metrics by a considerable margin, therefore performing the best in this real application.

To make the results more intuitive, we further provide a visual comparison on the aforementioned dark face detection in Fig. 8. We deliberately select a challenging sample that is extremely dark and contains 28 dark faces according to the official label provided. As illustrated in Fig. 8(a), the raw detection on this challenging sample yields very poor performance with only 1 face detected and 27 missed. In comparison, the enhanced detection generally performs much better, as can be observed from the rest of the figure. Nevertheless, upon deeper and closer inspection, we still find some notable differences among the enhanced results, leading to varied detection performance. For instance, the detection results of KinD and URetinex-Net are inadequate because their enhanced images are too blurry to recognize faces. Although the images enhanced by RRDNet and EnlightenGAN are clear, they are not bright enough in comparison to the others. While Zero-DCE, RUAS, SCI, and PairLIE generate visually decent images, the corresponding detection on them indicates that they still fail to recognize some faces (particularly the darker and smaller ones) to a certain extent. In contrast, our method provides the superior visual results by not only illuminating the faces in extremely dark regions but also preserving the details in well-lit areas. Consequently, it achieves the best detection performance with 23 faces detected out of 28, demonstrating the effectiveness and practicality of our proposed architecture and the great benefits of non-uniform enhancement.

V. Conclusion

In this paper, we introduce a novel framework called residual quotient learning for zero-reference low-light image enhancement. Unlike existing Retinex-related structures, our framework reformulates the low-light enhancement task as adaptively estimating the latent quotient with reference to the original low-light input using a residual learning manner. This makes our framework relatively simple yet physically explainable in terms of Retinex theory. Building upon this framework, we propose ResQ-Net, a lightweight and effective network with enhanced capabilities for modeling non-uniform illumination. Our carefully designed framework supports zero-reference training, significantly improving its generalization and adaptability. Extensive experimental results validate the effectiveness of the proposed residual quotient learning framework and network structure. Our trained ResQ-Net surpasses many state-of-the-art methods both qualitatively and quantitatively.

In future work, we plan to extend the residual quotient learning framework to other image restoration tasks with appropriate modifications. Furthermore, we will explore more advanced modules to better address intensive noise and mitigate color distortions, both of which remain significant challenges in extreme low-light scenarios.

Acknowledgment

This work was conducted when Chao Xie was a Visiting Scholar with Nanyang Technological University, Singapore. Part of this research was carried out with the Centre for Advanced Robotics Technology INnovation (CARTIN) Laboratory, Nanyang Technological University. The CARTIN Laboratory is supported by the National Research Foundation, Singapore, under its Medium-Sized Centre Funding Scheme (CARTIN). The opinions, findings, and conclusions expressed in this material are those of the authors and do not reflect the views of the National Research Foundation, Singapore.

References

[1] W. Hu, Y. Yang, and H. Hu, "Pseudo label association and prototype-based invariant learning for semi-supervised NIR-VIS face recognition," *IEEE Trans. Image Process.*, vol. 33, pp. 1448–1463, 2024.

[2] C. Xie, W. Zeng, and X. Lu, "Fast single-image super-resolution via deep network with component learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3473–3486, Dec. 2019.

[3] C. Xie, H. Zhu, and Y. Fei, "Deep coordinate attention network for single image super-resolution," *IET Image Process.*, vol. 16, no. 1, pp. 273–284, Jan. 2022.

[4] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1921–1935, Sep. 2009.

[5] S. Wang, J. Zheng, H. 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.

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

[7] 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.

[8] Z. Ying, G. Li, and W. Gao, "A bio-inspired multi-exposure fusion framework for low-light image enhancement," 2017, *arXiv:1711.00591*.

[9] 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.

[10] Y. Ren, Z. Ying, T. H. Li, and G. Li, "LECARM: Low-light image enhancement using the camera response model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 968–981, Apr. 2019.

[11] E. Land and J. McCann, "Lightness and Retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, p. 1–11, 1971.

[12] E. H. Land, "The Retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, Dec. 1977.

[13] S. Zheng, Y. Ma, J. Pan, C. Lu, and G. Gupta, "Low-light image and video enhancement: A comprehensive survey and beyond," 2022, *arXiv:2212.10772*.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] K. G. Lore, A. Akintayo, and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognit.*, vol. 61, pp. 650–662, Jan. 2017.

[16] F. Lv, F. Lu, J. Wu, and C. Lim, "MBLLEN: Low-light image/video enhancement using CNNs," in *Proc. Brit. Mach. Vis. Conf.*, Jan. 2018, vol. 220, no. 1, p. 4.

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

[18] 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.

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

[20] J. Li, J. Li, F. Fang, F. Li, and G. Zhang, "Luminance-aware pyramid network for low-light image enhancement," *IEEE Trans. Multimedia*, vol. 23, pp. 3153–3165, 2021.
[21] S. Lim and W. Kim, "DSLR: Deep stacked Laplacian restorer for low-light image enhancement," *IEEE Trans. Multimedia*, vol. 23, pp. 4272–4284, 2020.
[22] C. Li et al., "Low-light image and video enhancement using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9396–9416, Dec. 2022.
[23] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex decomposition for low-light enhancement," 2018, *arXiv:1808.04560*.
[24] 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.
[25] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3291–3300.
[26] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proc. 27th ACM Int. Conf. Multimedia (ACM MM)*, Oct. 2019, pp. 1632–1640.
[27] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1013–1037, Apr. 2021.
[28] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, and J. Jia, "Underexposed photo enhancement using deep illumination estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6849–6857.
[29] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "URetinex-Net: Retinex-based deep unfolding network for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5901–5910.
[30] J. Hai et al., "R2RNet: Low-light image enhancement via real-low to real-normal network," *J. Vis. Commun. Image Represent.*, vol. 90, Feb. 2023, Art. no. 103712.
[31] Y. Jiang et al., "EnlightenGAN: Deep light enhancement without paired supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.
[32] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
[33] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.
[34] J. Hoffman et al., "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1989–1998.
[35] Z. Ni, W. Yang, S. Wang, L. Ma, and S. Kwong, "Towards unsupervised deep image enhancement with generative adversarial network," *IEEE Trans. Image Process.*, vol. 29, pp. 9140–9151, 2020.
[36] L. Zhang, L. Zhang, X. Liu, Y. Shen, S. Zhang, and S. Zhao, "Zero-shot restoration of back-lit images using deep internal learning," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1623–1631.
[37] A. Zhu, L. Zhang, Y. Shen, Y. Ma, S. Zhao, and Y. Zhou, "Zero-shot restoration of underexposed images via robust Retinex decomposition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[38] C. Guo et al., "Zero-reference deep curve estimation for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 1780–1789.
[39] C. Li, C. Guo, and C. C. Loy, "Learning to enhance low-light image via zero-reference deep curve estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4225–4238, Aug. 2022.
[40] R. Liu, L. Ma, J. Zhang, X. Fan, and Z. Luo, "Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10561–10570.
[41] L. Ma, T. Ma, R. Liu, X. Fan, and Z. Luo, "Toward fast, flexible, and robust low-light image enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5637–5646.
[42] Z. Fu, Y. Yang, X. Tu, Y. Huang, X. Ding, and K.-K. Ma, "Learning a simple low-light image enhancer from paired low-light instances," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 22252–22261.
[43] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "Band representation-based semi-supervised low-light image enhancement: Bridging the gap between signal fidelity and perceptual quality," *IEEE Trans. Image Process.*, vol. 30, pp. 3461–3473, 2021.
[44] J. Li, X. Feng, and Z. Hua, "Low-light image enhancement via progressive-recursive network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4227–4240, Nov. 2021.
[45] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.
[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2018, pp. 7132–7141.
[47] Q. Fan, J. Yang, D. Wipf, B. Chen, and X. Tong, "Image smoothing via unsupervised learning," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 1–14, Dec. 2018.
[48] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Inst.*, vol. 310, no. 1, pp. 1–26, Jul. 1980.
[49] 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.
[50] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[51] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. workshops*, Jan. 2019, pp. 63–79.
[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[53] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[54] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 97–104.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[56] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Trans. Image Process.*, vol. 30, pp. 2072–2086, 2021.
[57] C. Lee, C. Lee, Y.-Y. Lee, and C.-S. Kim, "Power-constrained contrast enhancement for emissive displays based on histogram equalization," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 80–93, Jan. 2012.
[58] 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.
[59] W. Zhou, 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.
[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[61] 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, Apr. 2012.
[62] M. Behzadian, S. K. Otaghsara, M. Yazdani, and J. Ignatius, "A state-of the-art survey of TOPSIS applications," *Expert Syst. Appl.*, vol. 39, no. 17, pp. 13051–13069, Dec. 2012.
[63] W. Yang et al., "Advancing image understanding in poor visibility environments: A collective benchmark study," *IEEE Trans. Image Process.*, vol. 29, pp. 5737–5752, 2020.
[64] J. Li et al., "DSFD: Dual shot face detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5060–5069.
[65] S. Yang, P. Luo, C. C. Loy, and X. Tang, "WIDER FACE: A face detection benchmark," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5525–5533.

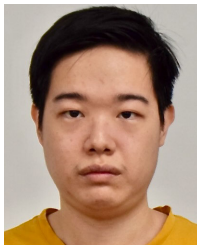
Chao Xie (Member, IEEE) received the Ph.D. degree from Southeast University, Nanjing, China, in 2018. From July 2023 to July 2024, he was a Visiting Scholar with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an Associate Professor with the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing. His current research interests include image processing and deep learning.

Linfeng Fei is currently pursuing the bachelor's degree with the College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing, China. His current research interests include image enhancement and deep learning.

Huanjie Tao received the M.S. degree in mathematics and information science from Capital Normal University, Beijing, China, in 2016, and the Ph.D. degree in control science and engineering from Southeast University, Nanjing, China, in 2020. He was a Visiting Ph.D. Student with the School of Computer Science and Engineering, Nanyang Technological University, from September 2018 to September 2019. He is currently employed as an Associate Professor with the School of Computer Science, Northwestern Polytechnical University, Xi'an, China. His current research interests include image processing, deep learning, and large language models.

Yaocong Hu received the B.S. degree in automation from Anhui Polytechnic University, Wuhu, China, in 2014, the M.S. degree in pattern recognition and intelligent system from Anhui University, Hefei, China, in 2017, and the Ph.D. degree in pattern recognition and intelligent system from Southeast University, Nanjing, China, in 2021. He is currently a Lecturer with Anhui Polytechnic University. His current research interests include image processing and deep learning.

Wei Zhou received the M.E. degree from Beijing University of Posts and Telecommunications, China, in 2020. He is currently pursuing the Ph.D. degree with the School of Electronics and Information Technology, Sun Yat-sen University, China. His main research interests include computer vision, pattern recognition, and deep learning, in particular focusing on multi-label image classification, image captioning, and weakly-supervised temporal action localization.



Jiun Tian Hoe (Student Member, IEEE) received the B.CS. degree from the Faculty of Computer Science and Information Technology, Universiti Malaya, Malaysia, in 2022. He is currently pursuing the Ph.D. degree with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, under the supervision of Prof. Yap-Peng Tan and Prof. Xudong Jiang. His research focuses on deep learning, computer vision, and image generation and their applications.

Weipeng Hu received the Ph.D. degree in electronics and information technology from Sun Yat-sen University, Guangzhou, China, in 2022. He is currently a Research Fellow with the Centre for Advanced Robotics Technology Innovation Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His current research interests include image and video synthesis, human–robot interaction, heterogeneous face recognition, and cross-domain person ReID.

Yap-Peng Tan (Fellow, IEEE) received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1993, and the M.A. and Ph.D. degrees in electrical engineering from Princeton University, Princeton, NJ, USA, in 1995 and 1997, respectively. From 1997 to 1999, he was with Intel Corporation, Chandler, Arizona, and Sharp Laboratories of America, Camas, Washington, USA. In November 1999, he joined Nanyang Technological University (NTU), Singapore, where he is currently a Professor and the Associate Vice President (Lifelong Learning Postgraduate Programmes by Coursework). His current research interests include image and video processing, content-based multimedia analysis, computer vision, pattern recognition, machine learning, human behavior analysis, and data analytics. He had served as an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE SIGNAL PROCESSING LETTERS, and IEEE ACCESS; and an Editorial Board Member of *EURASIP Journal on Advances in Signal Processing* and *EURASIP Journal on Image and Video Processing*.