TBEFN: A Two-branch Exposure-fusion Network for Low-light Image Enhancement

Kun Lu hzmylys@gmail.com
Lihong Zhang† lhzhang@sxu.edu.cn
Shanxi University

Fig. 1. Comparison of the low-light input (left of the red line) and our enhanced images(right). Horizontally symmetrical images are used for fairness. Details in low-light images tend to lost in noise and dark zone, while the proposed method yield a much more pleasant result.

Abstract

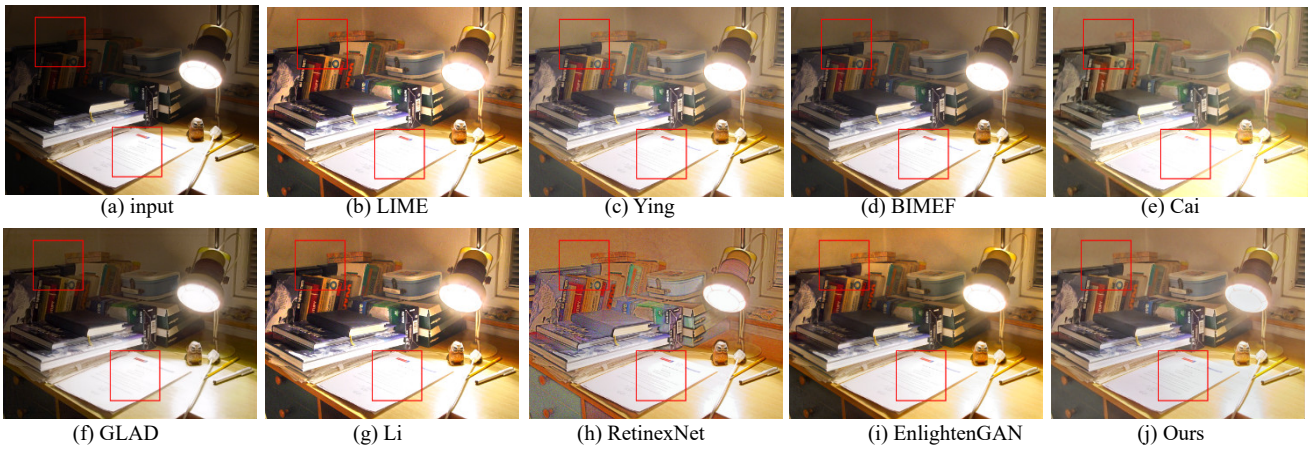
Most of the existing low-light image enhancement methods are performed with one fixed branch, where images of low exposition levels are tried to be mapped into a well-illuminated ones. But these fixed framework is not robust enough and lacks some flexibility when faced with varied scenes and exposition levels. In this paper, we propose a two-branch exposition fusion network (**TBEFN**) to deal with the problem. First, the input images are enhanced in the two-branch enhancing module where appropriate transformation function from -1E and -2E exposition levels are estimated with or without pre-denosing procedure according the targeted luminance. A self-adaptive judging module would produce attention maps for the two candidates and have them fused. To better fight against detail loss and unintended noise, we use a fine-tune module to produce the ultimately enhanced images. In our experiment, three benchmarks are used for quantitative evaluation and six for visual evaluation, where the proposed one outperforms many of the existing state-of-the-art ones, showing great robustness and enhancing ability of this work. Moreover, discussions about the loss function and enhancing theory are provided for future study.

I. Introduction

Human beings tried to record the world with images at an early age. In most cases, however, a well-illuminated images with clear details and vivid colors are often difficult to be obtained. This happens when the surrounding environments are relatively dark (e.g., in the night or with limited illumination) or simply because the cameras are not adjusted properly (e.g., an inappropriate ISO, EV or Exposure time). Unlike scenes with multi-exposure samples [1, 2], the real-world images are usually taken once in a given location, which makes the enhancement task more difficult. Moreover, the enhancement of low-light images is generally more challenging compared with some other low-level image enhancing tasks, due to their accompanying extremely dark zones, unexpected noise and blurred details in particular. See left regions in Figure 1 for more details.

To restore these low-light images whilst remaining the key semantic information, earlier methods try to redistribute the values of the pixels using varied nonlinear mapping strategies, where the most representative methods are histogram equalization [3-5] and gamma correlation [6]. Another kind of classic method is based on varied physical priors, such as Retinex theory [7-9], camera response system [10], dark channel prior [11, 12] and absorption light scattering model [13]. Though classic methods are designed for a wide range of low-light conditions, they are sensitive to the elaborately selected parameters, but the optimization process requires large amounts of knowledge and experience. Directly applying these methods for varied illumination conditions tend to fail since they are not powerful enough to restore details in an image (Figure 2(d)), or over-enhancement (Figure 2(b) and (g)). Since classic methods

† Corresponding author.



(a) input (b) LIME (c) Ying (d) BIMEF (e) Cai

(f) GLAD (g) Li (h) RetinexNet (i) EnlightenGAN (j) Ours

Fig. 2. Performance on an image in MEF dataset. Existing methods are more vulnerable to extremely dark and bright regions, where severe noise and detail loss are observed.

are not specifically designed in the denosing task, visible heavy noise is observed in dark regions. Recently, thanks to the progress of computational resources and publication of massive databases, data-driven method has been studied extensively, where CNN-based and GAN-based methods [14-17] have shown great potential to generate a satisfactory result. However the fact that elaborately designing and training a neural network has always attracted researchers' attention, there still exists a lot room to improve. In Figure 2 (e, f, and h), for instance, the existing methods trained on a specific dataset tend to show their favor to a certain range of illumination conditions and scenes, and thus is limited by its generalization ability. Though GANs are trained on unpaired datasets, they sometimes yield unexpected colors which might looks unnatural or over-saturated (Figure 2(i) and 6(b)). In addition, a direct image-to-image transformation is relatively more difficult than that combined with physical priors [18], and this turns to be more challenging if noise under low-light conditions is taken into consideration.

In order to better overcome the problems discussed above, in this paper, we present a novel fusion structure based on CNNs. The primary point of our method is to tackle with the problem of both noise suppression and robustness on different illumination conditions, which is widely discussed but barely handled properly in the existing methods. The structure of our proposed method is inspired by Retinex theory and multi-fusion strategy, since more enhancing trails tend to yield more candidates, and so better results. Since -1E and -2E illumination conditions are the mostly common situations, the first stage is composed of a two-branch basic enhancing framework where two independent basic enhancing modules are used to estimate the illumination transfer function from -1E and -2E exposure levels respectively. Considering the heavy noise in -2E environment, a specifically designed pre-denoising model is further introduced. By this approach, we get two candidates with varied illumination levels and styles, which provides more chances for the model to generate a satisfactory result. These images are then fed into the second stage where a judging module is trained to fuse the inputs with a simple yet effective attention mechanism. Moreover, a fine-tune module is introduced as a fine-tuner to perform the final post-processing denosing procedure and detail composition before images are qualified to be outputted as the final result.

The contribution of our work is three-fold:

1) We propose a two-stage fusion network for low-light image enhancement where the two basic enhancing branches are utilized to generate the final result, and thus generalization ability to different illumination conditions compared with existing methods. Moreover, the pre-denoising and post-denoising procedures are performed along with our enhancing structure, which is shown to be effective in noise suppression when tested on challenging datesets.
2) We propose to use SSIM loss to compensate for the structural blindness of *l1* and *l2* loss, and distance for feature maps produced by VGG could be another complementary metric to fight against dataset defects. Moreover, discussions on methods based on Retinex theory are given in this paper, where we theoretically believe that it is more appropriate to have the luminance re-estimated than directly regarding the reflection as the final enhanced result.
3) We performed a comprehensive evaluation for most of the existing state-of-the-art methods on three paired datasets and six commonly used unpaired dataset. Further, the experiment shows that our proposed method achieves a competitive result.

II. Related Work

A. Retinex-based Image Reconstruction

Image decomposing is an intuitive yet effective method in the low-level image processing area. It is most commonly used when certain properties or components are particularly concerned, and could be further used to facilitate a given task, such as dehazing, denosing, super resolution, and so on [19-21]. The most representative and commonly used decomposition strategy in the low-light image enhancement area is the Retinex theory [7], which aims at decomposing an image into its illumination component L and reflection R. Namely

$$I(x, y) = L(x, y) \circ R(x, y) \quad (1)$$

where $\circ$ is the element-wise product, L is basically assumed to be piecewise smooth and R is piecewise constant, representing the reflectance property of the observed object.

As seen from Eq. 1, early attempts believe that once the reflection component L is estimated, the reflection component R, which is believed to be the unpolluted image according the visual system, would be then figured out using a basic division operation i.e. $R=I/L$. As such, some Retienx-based methods are dedicated to find an accurate estimation of the luminance component L. In the very early attempts of SSR and MSR [8, 9], one or more Gaussian filters are used to generate a smooth estimation of the illuminance, where the estimating process is actually an weighted average of neighboring pixels. Using a coarse-to-fine strategy, the authors in [22] first estimate the luminance component by initiating the illumination with the max RGB values for each pixel, and then impose a structure prior to have it refined. Fast though these methods are, their processed images tend to be unnatural over-enhanced, as many recent observations have found [23-25]. This is because illumination itself is still an important factor in the human vision perception system, which would be discussed in the next section.

Instead of removing the luminance component, another category of Retinex-based methods estimates the illumination and reflectance simultaneously, which are then both utilized to generate the final enhanced image. In this manner, the decomposition process is generally constrained with more strict priors on the reflectance and illumination [15, 16], and the whole process is relatively computational intensive. Recently, [21] added a color term in Eq. 1 and obtained a good result by decomposing the image into three components, which is also a variant of the classic Retinex-based decomposition method.

B. Methods for low-light image enhancement

Based on varied enhancing mechanism, methods for low-light image enhancement could roughly be divided into three categories: direct enhancement, prior-based and data-driven schemes.

Directly Enhancing Approach. The key idea of these methods is to increase the contrast and dynamic range of a given image. One of the most commonly used technique is histogram equalization (HE) [3], which aims to artificially force the pixels distributed evenly in the histogram. This method, along with many of its variants, who further exploit additional information such as inter-pixel contextual [4] and gray-level differences [5], however, are limited in their performance since the assumption that an evenly distributed histogram is not ideally matched in reality, and the histogram of an input image is barely considered [13], which means the enhancing mechanism itself is of some defects and thus limited in generalization ability. As another direct approach, Gamma correction (GC) [6] is intended to expand the dynamic range of an image with an exponential function for each individual pixel. This, however, inevitably ignores the correlation between a pixel's neighbors and that in different channels, thus having a built-in tendency towards artifacts and amplified noise.

Prior-based Approach. The second category of methods is built on varied decomposition priors. Retinex theory, as mention above, is the most widely adopted one. Following the early attempts, [26] is dedicated to preserve the naturalness by using a bright-pass filter and bi-log transformation algorithms, where a new assessment metric is introduced as well. In [23], a method based on weighted multi illumination fusion is studied to balance image details, contrast and naturalness. With detailed observation of the log-transformation on Retinex formula Eq. 1, [24] introduced an exponential term to compensate the high luminance component in log-transformation where high-light parts are buried in irrelevant low-light ones. These methods, though with a carefully designed architecture, are vulnerable to extremely dark zones with unexpected noise, and tend to cause some color distortion. Aside from classic enhancing methods, many of the data-driven approaches also adopt Retinex theory as the theoretical baseline. For example, [27] believes that SSR/MSR-based method is equivalent of using different Gaussian kernels. Recently, with the accessibility of large scale paired dataset, in [14, 16, 18, 28] the authors use convolution neural networks to perform the estimation and reconstruction of both reflection and illumination. However, these networks are mainly trained on a specific dataset or exposition ratio, and could not adjust to more complex scenarios, Still, combining with various loss function, the training process is much labor-consuming.

Apart from the widely-used Retinex-based prior, many other decomposition priors are exploited to accomplish effective low-light image enhancement. Some researchers find that inversed low-light images much resemble those taken in the haze, and thus methods primarily designed for dehazing task is also applied here [11]. Very recently, based on carefully observation on the atmosphere light, [13] proposed an absorption light scattering model which jointly enhances an low-light image and alleviates noise. However, this model is somewhat limited to its performance and is extremely computational-intensive.

Data-driven Approach. Thanks to the publication of available datasets, data-driven approaches are now undergoing a huge development. It is worth mentioning that some of the works in this category also fall into the decomposition-based ones, as discussed above. Apart from the most commonly used reconstruction loss, which is usually $l1$ norm or SSIM loss since a smooth solution yield by $l2$ norm could hardly restore an image with a wide dynamic range [22, 29], constrains on reflection and illumination are also imposed according to the priors. Very recently, [18] added a color loss by comparing the distance on RGB angle and obtained a satisfactory result.

The other data-driven approaches, roughly speaking, are purely data-driven. [15, 30] trained a network directly by learning an image-to-image mapping function. In [31], a deep fusion network is proposed with dark-area-based mask to facilitate the learning of regional dark areas. [32] built a dataset consisting of images taken through different devices, and a CNN based neural network is further proposed. In their work, a composite perceptual error function combining color, texture, content and total variation loss is used for training. However, this network is designed for image transforming between different devices, but not typically for images under different lighting conditions, thus its performance is severely limited to mediate level low-light conditions. In [33], the authors built a multi-exposure dataset and further trained a step-enhancing network that predicts images under varied expositions from a well-

illuminated one. Since this network was trained on several sub-nets separately, the consistency between neighboring exposition levels are contents are not studied, resulting in inaccurate exposure and lost in contextures. Moreover, an off-the-shelf tone-mapping technique [34] is used to complete the reconstruction task. Very recently, [25] proposed a two-steam enhance scheme to describe content features and RNN-based edge details respectively, where a discrimination loss is adopted. Note that since this work is studied under artificially degraded training set, and no specific denoising technique is adopted, it has a relatively poor performance under real scenarios, where obvious color distortion is observed. To overcome the difficulty of preparing well-prepared datasets, [35] further proposed an GAN-based unpaired image enhancement where the loss function is much similar to their previous work [32]. Very recently, [17] proposed a simply U-net [36] structure with self-regularized loss function and attention mechanism and yielded a good result. Apart from methods dealing with commonly-used pictures, [29] also proposed an enhancing baseline for data in raw image format using its corresponding dataset, which is undoubtedly somewhat limited by its usage.

III. METHODOLOGY

A. Discussion about Retinex-based Methods

It is interesting to find that different Retienx-based methods, whether the luminance component is considered in the reconstruction process, both achieve a satisfactory result. This, however, poses a question to think about why. First, considering the early attempts of Retinex theory, if the illuminance is directly removed from a low-light image, we'll get a well-illuminated one, say, the image of the reflectance property of the observed object. Then, here comes the dilemma that: what is the illumination of the enhanced one, since the enhanced one itself is also a picture, and is further observed by us as well. One possible explanation is that, if the illumination component is equal to an all-ones matrix, then the enhanced one equals the reflectance. This, however is somewhat a little confusing to the prior that the luminance component is partially contingent on the structure of objects, which varies among different structures in the image. As such, we may suppose that directly having the illumination removed actually conflicts with the original Retinex theory itself, which is dedicated to shown how images are formed. In other words, Retinex theory describes that the perception of an image is composed of its reflectance and luminance componence, but not merely the reflectance. For example, let I denote a well-illuminated image, and I' the degraded low-light one, we have:

$$\begin{cases} I = R \circ L \\ I' = R \circ L' \end{cases} \tag{2}$$

where L' is the luminance under low light conditions. Our task is to estimate the well-illuminated image I using the observed low-light one, instead of estimating the reflection R, which is the natural reflection property of an object, but not the expected imaged itself.

As such, the enhancement task is express as

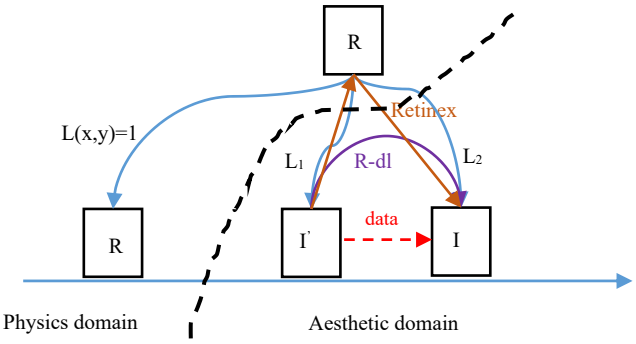


Fig. 3. Enhancing routines for Reinex-based classic methods, Retinex-based data-driven methods and purely data-driven approaches. We shown that illumination is a key factor to compose a well-illuminated image.

$$I = I' \circ \frac{L}{L'} \tag{3}$$

Specifically, when L is believed to be an all-ones matrix, *equ(3)* is the same with the methods that directly remove the illumination in log field. Since the true L is somewhat correlated with the structure, the enhanced image is clearly not that satisfactory, just as many research have observed. To distinguish from the early observations, data-driven methods actually model the transformation $1/L^*$ between the enhanced images and the corresponding observed scenario. Note that in these cases, $1/L^*$ is an estimation of L/L', but not the relationship between the images and natural properties with various priors and strict constraint on the decomposition process, which is $I=RL$. (see Figure 3 for more details)

In the broadest sense, when R and L are separately estimated using the observed I', the enhancement would be performed based on the basic Retinex theory, which makes it possible to transfer the original task from an image-to-image transformation to transformations between its components and thus facilitating a precise enhancement. Nonetheless, the computational cost tends to increase due to the task of both restricting the decomposing and the reconstructing process.

It is worth mentioning that, we argue that these two enhancing methods both work theoretically, and the major difference is the estimation difficulty and approach: the prior one models the transformation function, while the later one models the strict decomposition function. The reason why these methods could both achieve satisfactory results is explained by the imaging mechanism since an enhanced image is obtained by multiplication. The direct estimation of L^* is L'/L, which is exactly expressed as L' multiplexing L^{-1} in the second method. However, note that early attempts treating L^* as L' tend to fail due to the fact that L is not equal to 1; still, they fail to model the transmission function which is difficult to get obtained without sufficient data and high-performance computational resources. To express the relationship between these two methods, we have

$$I = I' \circ \frac{1}{\hat{L}} = I' \circ (\frac{\hat{L}}{L'}) = I' \circ \hat{L}^{-1} \circ \hat{L} = \hat{R} \circ \hat{L} \tag{4}$$

where the first and the last equation represent the estimation of transformation function and Retinex theory respectively. Based on the above discussion, we now describe the relationship of varied enhancing methods in Figure 3, where we demonstrate that process of low-light image enhancement intrinsically must start and end on the existing points in the human reception field, while the explicit decomposition into reflection and illumination is a prior-guided strategy with solid physical explanation.

B. Motivation

In our previous discussion, the observed degraded image is intrinsically considered to be noise-free, which is rarely satisfied considering the status quo, and that is one of the major culprit for why many of the existing methods could not yield their expected result in real-world scenarios. Since the noise term is usually considered to be additive, and both to simplify our discussion, *equ(2-3)* could be rewritten as:

$$\begin{cases} I = R \circ L \\ I' = R \circ L' + E \\ I = (I' - E) \circ \dfrac{L}{L'} \end{cases} \tag{5}$$

equ(4) indicates that an appropriate pre-denoising procedure would help to eliminate the noise in an observed image. Thus the enhancement could be divided into two sub-problems where the noise E and transformation function L/L' are estimated sequentially.

To resolve the noise problem, the authors in [15, 22, 30] tend to some off-the-self denoising techniques, such as BM3D [37] and pre-trained neural networks. However, researchers have found that noise in low-light images methods tend to take on more complicated patterns [38, 39], which means that many of the off-the-shelf denoising techniques might not achieve our expectation in extremely dark regions compared with normal luminance, and further a specifically designed denoising module ought to be introduced.

However, a pre-denoising procedure consist with Eq. 4 is relatively difficult to be trained in a deep neural network, and the denoising process under low-light conditions itself is a challenging task, where elaborately paired images or solid theoretical priors are rarely found comparing with that under well-illuminated conditions. To deal with this, the proposed enhancing framework is designed to be a two-stage one, where data-driven noise suppression is carried out both at the input and output ports. By turning a challenging denoising task into two smaller ones, our experiments have further shown the superb denoising ability compared with other stare-of-the art methods.

Our another motivation comes from the aesthetic process of human beings that a painter is more sophisticated in one of the given scenes (i.e, animals, buildings, plants), but when asked to perform on their unfamiliar scenes, more trails would be carried out to obtain a relatively satisfactory result. That's to say, it is simpler to train a judge to fuse different enhancing samples than asking one method to yield that an excellent generalization ability for varied illumination conditions typically under the discussion of this paper.

Based on these discussion, a two-steam fusion network is introduced to enhance images under exposure level of -1E and -2E typically, which are the most common-seen low-light situations. Then, these enhanced images are fed into an attention-based fusion module where a judge is trained to have them fused with the original low-light image. The whole structure of the proposed method is shown in Figure 4 and Figure 5.

C. Network Structure

1) Basic enhancing module: We use the basic enhancing module to estimate the transfer function between different lamination levels, which is

$$I_{\substack{ex \\ ex \in \{-1E,-2E\}}} = I' \circ \mathcal{F}_{ex}(I') \tag{6}$$

Empirically, a basic enhancing module should be equipped with enough transferring ability whilst remaining the details and global style of the input image. As observed by many researchers,[14, 15, 29, 40], the U-net structure showed great performance in many of the image-to-image transforming tasks. The encoder-decoder structure and skip connection into deeper

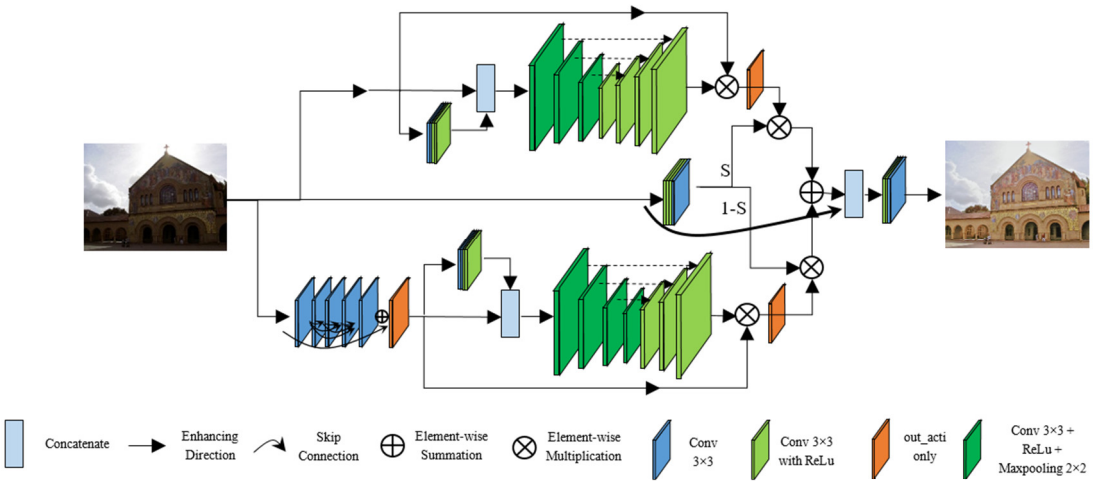


Fig. 4. Structure of the proposed model

(a) (b) (c) (d) (e) (f) (g)

Fig. 5. Enhancing precedures of the proposed model. (a) Input, (b) Results of the -1E branch, (c) Denoised output of -2E, (d) Results of the -2E branch, (e) Attention map, (f) Fused results of two branches, (g) Final output from the fine-tune module. The attention map mainly concentrates on details and high-light regions of -1E branch, and noise is suppressed through the -2E branch. A fine-tune module is adopted to compensate for the content loss of the fused image (lost details, luminance, and noises). One can observe the obvious noise suppression between (f) and (g) from the zoomed-in patch.

layers enable this structure to learn both the multi-scale features and latent transfer function between varied illumination conditions. However, since the channels of feature maps increase exponentially as the layers come close in the middle, the computing cost and difficulty of back-propagation increase accordingly. Since we are mainly concentrated on the multi-scale encoder-decoder structure, a simplified one with channels increasing from 16 to 128 is used in this paper, whose parameters are significantly fewer than that in the existing methods. This set of parameter is adopted since an increase on the parameters did not seen a sharp improvement in our experiments, but adding more computing cost. Note that four convolution layers are carried out first on the input image to produce its detailed map, which is then concatenated with the input low-light image before they are fed into the basic enhancing module.

2) Pre-donoising module: Since the noise is basically considered to be addictive, our denoising module is thus constructed on dense skip links with element-wise addition performed consequently. A total of five convolution layers are performed with a kernel size of 3×3, with each of the last four layers added with all the output results from the former layers. A dense connection in this manner was proved to improve the learning ability in deep networks [41]. Moreover, to fully utilize the layer-to-layer transferring information, no activation function is performed after the convolution layer except for the last one, where a modified ReLU function is used to facilitate the training process, which is expressed as

$$out\_acti(x) = relu(x) - relu(x-1) \tag{7}$$

Compared with sigmoid activation, the proposed activation could both map the input features into a range of [0,1] and keep a relative high derivation in the back propagation process. Note that we view the estimated noise as one fallen into the range of $(-\infty, +\infty)$, which has a more complex pattern and is more suitable in the low-light situations.

3) Judging module: In this module, a judge need to be trained so as to distinguish both strengths and defects of the candidate images enhanced by our two-stream network where -1E and -2E exposure are separately enhanced. Analogous to human perception, our judge first judges the amount of enhancement required based on its own observation of the input original image, and then fuses these two input candidates. In this paper, we use a similar four convolution layers to produce the 1-D attention map S on the -1E enhanced image, and the complimentary component 1-S for the -2E one. The fused image could be expressed as

$$\underset{c\in\{R,G,B\}}{I_{fu}^{c}} = I_{-1E} \circ S^{c} + I_{-2E}(1 - S^{c}) \tag{8}$$

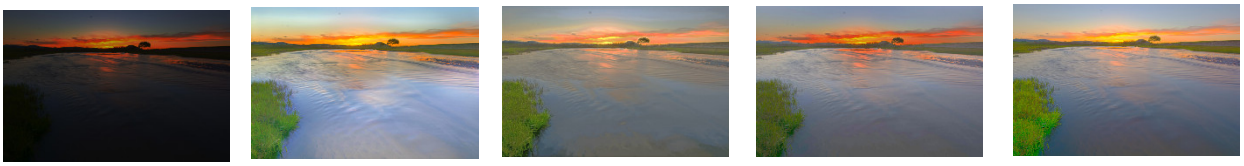
Simple though the method is, the main drawback is that our enhanced -1E and -2E images are produced separately, where some key features might get lost during the enhancement, and noise tend to be amplified using such a direct metric. To tackle with this problem, the fused image I_{fu} is fed into a refinement convolution module with its low-light input concatenated. The final enhanced image $\hat{I}$ is expressed as

$$\hat{I} = \mathcal{F}_{ref}(concat(I', I_{fu})) \tag{9}$$

Note that the refinement module performs post-denoising and detail refinement simultaneously.

D. Loss Function

1) SSIM loss: In our experiment, we found that *l2* loss tends to fail when directly used as the loss function. This is probably explained by the unique property of the enhancing task. Since we are dedicated to enlarge the dynamic range and contrast of a low-light image, it is more suitable to yield an image with sparse distribution of pixels. Though *l1* loss could compensate that for some extent, it still ignores the structure of an image and might cause severe halos on the tested images, which means that *l1* and *l2* loss are structure-blind. Figure 6 shows one case where a picture with high psnr is proved to be visually unacceptable with severe structural halos. Though direct pixel-to-pixel metric (*l1*-distance and *l2*-distance) performs well in Figure 6(b) and (d), it fails in (c). Since SSIM jointly evaluates the luminance, contrast and structure, it could better satisfy our primary goal compared with that trained on *l1* or *l2* loss, which is consistent with the observation from many other experiments [14, 29]. Following



(a) input (b) EnlightenGAN (c) Cai (d) Ying (e) ground truth

Fig. 6. An example of structure blindness of pixel-to-pixel evaluation metric. The PSNR/*l1*-distance (per pixel) /SSIM/vgg-distance of three methods are (b): 16.87/32.65/0.84/0.66; (c): 23.94/13.46/0.89/0.60; (d): 21.63/16.30/0.91/0.39. Here we see that pixel-to-pixel metric contradicts with that of structural metric, but obviously the result of (d) is more appealing to human perception.

their path, we use the SSIM loss function instead, namely

$$\mathcal{L}_{SSIM} = 1 - SSIM(\hat{I}, I) \tag{10}$$

2) VGG loss: We use VGG loss as a complementary term for two reasons. First, It could be seen from [42] that as *l1* and *l2* loss ,SSIM is not that sensitive to changes of colors, which might cause a slight color distortion. Second, we found that the training and evaluation dataset from Cai *et al.* [14] is not perfectly matched. In other words, the ground truth still has certain color and structure artifacts and looks somewhat unnatural, which is shown in Figure 6(e) and Figure 7(1j). This is caused by the property of artificially synthesized images, since the ground truth is obtained by a combination of several off-the-shelf enhancing methods at that time. And thus image-level pixel-to-pixel loss functions would not perfectly depict the quality of our desired ones. Taking the two reasons into consideration, we adopted the output of conv_2_2 and conv_3_3 from a pre-trained VGG16 network to represent the style and color feature. MSE is used to measure the distance between ground truth and our enhanced results.

$$\mathcal{L}_{VGG} = \frac{1}{WHC} \sum \left\| \mathcal{F}_{VGG}(\hat{I}) - \mathcal{F}_{VGG}(I) \right\|^2 \tag{11}$$

Where *WHC* is the number of pixels in an image.

E. Implementation Details

We train the model with a three-stage strategy. First, the basic enhancing module is trained to convergence with L_{SSIM}. Since the images in -1E contain much less noise compared with the -2E ones, we only performed the denoising process for the -2E branch. Evaluations are executed every epoch, and the parameters producing the best results in -1E exposure level are selected to form the first enhancing branch. Second, the parameters with moderate score on both -1E and -2E are selected and frozen to train the denoising module in the second branch. After convergence, a joint training for branch-2 is carried out.

Fig. 7. Evaluation on the training dataset. These three images are from -1E exposure level, -2E exposure level of Cai's and LOL dataset respectively.

All the training in brach-2 are performed with L_{SSIM} only, and the best parameter set for -2E are selected to form the second branch. Finally, the judge module is trained with a similar strategy as the branch-2, the only difference is that the judge module is trained with a combination of L_{SSIM} and 0.1 times L_{VGG}, without any bias on scores of these two exposure levels. The model convergences quickly when fine-tuned in the last stage. All the training are performed with Adam optimizer [43] and Xavier initialization [44] by default in tensorflow, with a batch size of 10. The learning rate is set to be 1e-4 as a constant till the fine-tune stage where 1e-5 is adopted. Our experiments are done with a NVIDIA GTX 1080 GPU. The codes are available on https://github.com/lukun199/TBEFM.

IV. Experimental Results

A. Trainning Data

We adopt the dataset proposed in Cai [14], where 589 scenes are provided with captured image sequence of varied exposure levels and their reference ground truth synthesized with state-of-the-art algorithms. Since this dataset contains relative fewer indoor scenes, those images in LOL [16] are also used as a supplement. We follow the same route with the authors to separate the training set and evaluation set. The images in Cai's dataset are scaled to one fifths so as to alleviate the slight misalignment and 10 patches of 256×256 are randomly cropped for one image of each scene. For images in LOL dataset, we cropped 3 patched for each of the images. No data augmentation techniques are used. All the trainings were done with a mixture of 14,531 patches from Cai's dataset and 1,449 patches from LOL dataset.

B. Testing Data and Evaluation Metrics

Paired images with ground truth from Cai's and LOL dataset are used for quantitative evaluation. Following the routine of [16, 45], unpaired images from DICM, Fusion, LIME, MEF, NPE and VV [22, 26, 47, 48] datasets are adopted to perform the visual evaluation. These unpaired images are directly downloaded from testing images of [16], where tests for medium and well-illuminated images are not carried out. Note that due to the memory limitation., we scaled the images in Cai's dataset to one fifth of the original size and some large images that could not be processed with our machine are excluded from the statistics (Cai's model on VV dataset for example). The excluded images are given in the supplementary material.

For paired images, PSNR and SSIM are used to measure the distance between the enhanced ones and their referred ground truth. For both the paired and unpaired images, the most commonly non-reference metric NIQE [49] is directly adopted as the evaluation metric. Note that we compute PSNR and SSIM with python library skimage whose data range is set to be 255, and NIQE using MATLAB function niqe.

C. Results of Quantitative Evaluation

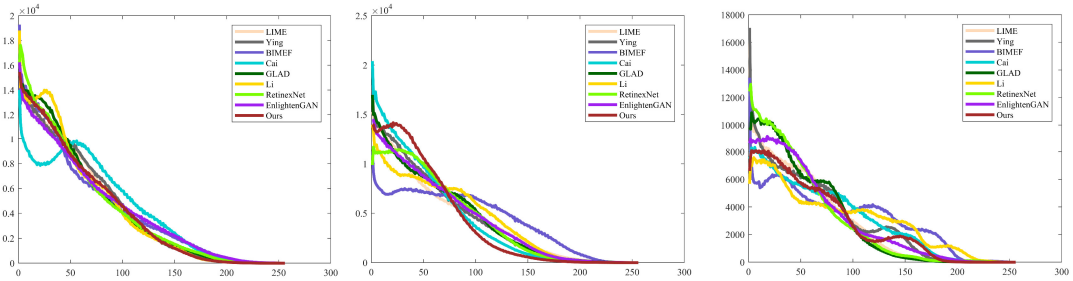
To make an objective and holistic comparison, we selected eight of the most representative and state-of-the-art methods including Retinex-based classic methods: LIME [22] and Li [46], prior based methods: BIMEF [45] and Ying [10], Retinex-based deep learning methods: RetinexNet [16] and Cai, [14] and purely data-driven methods: GLAD [15] and EnlightenGAN [17]. For fairness, all the codes are downloaded from the authors with recommended set of parameters. Both the input low-light image and their enhanced ones are provided.

1) Evaluation on Cai et al.'s Dataset: Since images in Cai's dataset are taken under varied exposure levels, we launched a comprehensive evaluation for the -1E and -2E images, which is the most commonly seen situation. The first two rows in Figure 7 show a representative image from the -1E illumination, where our method shows the strongest ability to restore colors and structural details. In contrast, GLAD, BIMEF and EnlightenGAN are not sensitive to the dark zone and lacks the ability to give a proper enhancement, while Cai, LIME, Li and RetinexNet showed their tendency towards over-enhancement and caused severe color distortion. Though Ying's method achieves a relative satisfactory results, it could not handle the amplified noise during the enhancement. In the darker situations (-2E), as shown in the third and fourth row, our proposed method is still powerful enough to reveal the details in extremely dark area whilst remaining robust to balance the light areas simultaneously.

2) Evaluation on LOL Dataset:

TABLE I
AVERAGE PSNR/SSIM/NIQE ON THREE TESTING DATASETS. THE BEST AND SECOND BEST RESULTS ARE MARKED IN BOLD AND UNDERLINE RESPECTIVELY.

Methods	Dataset								
	Cai-2E			Cai-1E			LOL		
	PSNR	SSIM	NIQE	PSNR	SSIM	NIQE	PSNR	SSIM	NIQE
SRIE[24]	13.68	0.665	3.1053	18.40	0.818	2.9682	11.86	0.495	7.5349
LIME[22]	17.22	0.733	3.4909	17.24	0.814	3.0725	16.92	0.504	8.7945
Ying [10]	19.60	<u>0.802</u>	3.2031	19.07	0.861	2.9080	17.20	0.623	8.0182
BIMEF[45]	15.91	0.748	3.1511	**20.45**	**0.867**	<u>2.8856</u>	13.88	0.595	7.6992
Cai[14]	**20.36**	0.796	3.3677	17.72	0.828	3.2021	16.16	<u>0.691</u>	<u>3.6176</u>
GLAD[15]	18.78	0.790	2.9794	19.71	0.851	2.9097	**19.72**	0.682	6.7970
Li [46]	17.04	0.724	3.6524	<u>20.11</u>	0.838	3.5025	13.88	0.664	3.9517
RetinexNet[16]	16.87	0.707	4.2616	16.67	0.760	3.8479	16.77	0.425	9.7281
EnlightenGAN[17]	16.98	0.757	**2.8987**	17.30	0.833	**2.8367**	<u>17.48</u>	0.652	4.8891
Ours TBEFN	<u>19.93</u>	**0.830**	<u>2.9693</u>	19.34	**0.863**	2.9496	17.31	**0.774**	**3.0439**



(a) (b) (c)

Fig. 8. Histogram of averaged pixel-to-pixel difference (per image) on three quantitative datasets. (a) -1E in Cai's, (b) -2E in Cai's, (c) LOL dataset. Best zoom in to see more details.

LOL dataset is composed of extremely dark images taken indoors. Severe noise and sporadic color information submerged in dark areas make the restoration work more challenging. One of the most representative images is shown in the last two rows in Figure 7. To better compare the performance of these methods, we select three important zoomed-in patches, as shown in the Figure 7. One can see that most of the tested methods are vulnerable of the change of application situations, and some even generate certain halos around high-contrast areas or could only produce globally insufficient enhancement. Moreover, images in Figure 7(3a-3h) are accompanied with much more noise than our enhanced result, which further demonstrates the effectiveness of our two-stage denoising strategy. However, our proposed method still lacks some capacity for color restoration, which is also shown in Figure 7(3i).

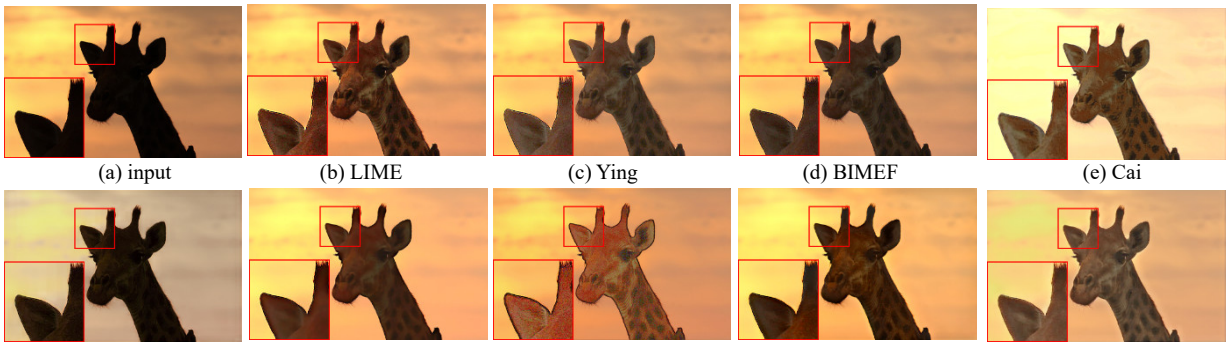
To better show the difference between the enhanced results and their ground truth, the pixel-level difference is calculated and then plotted into a histogram. More pixels fallen into a small difference value means a better enhancing performance. As shown in Figure 8, our method (shown in the brown line) achieved an excellent performance in a total of nine candidate approaches. Quantitative evaluations on these three datasets are given in Table I.

D. Results of Visual Evaluation

Evaluation on the non-reference dataset is a significant approach to verifying the generalization ability of the proposed model. The major challenging of this test is the blindness of the input images, which might be taken under all kinds of luminance and noise patterns. We evaluate the NIQE score of all six datasets and the averaged NIQE was recorded in Table II. It is not surprising that the proposed method outperforms many of the state-of-the-art methods since the two-branch and judging structure could provide more enhancing choice for the judging module, who would accordingly perform a proper fusion based on the recognized patterns of the input image. We find that our model behaves better in 1) a proper restoration for degraded colors, 2) adequate noise suppression capacity compared with other methods, 3) robust and of better generalization ability for

TABLE II
NIQE ON SIX UNPAIRED DATASET.
THE BEST AND SECOND BEST RESULTS ARE MARKED IN BOLD AND UNDERLINE RESPECTIVELY.

	DICM	Fusion	LIME	MEF	NPE	VV
LIME[22]	3.5186	3.3066	4.3466	4.3341	3.5921	2.2333
Ying[10]	3.3624	3.0187	3.9146	3.9162	3.3929	2.1453
BIMEF[45]	3.2659	2.9146	3.8069	3.8249	3.2818	**2.0621**
Cai[14]	3.1790	2.9431	3.8122	3.6654	<u>3.2489</u>	-
GLAD[15]	3.1399	3.0151	4.1209	3.6643	3.2921	<u>2.0904</u>
Li[46]	3.3186	3.7960	4.1080	4.1661	4.0486	-
RetinexNet[16]	4.7125	4.0583	5.1279	5.6314	4.4657	-
EnlightenGAN[17]	<u>2.7799</u>	<u>2.8211</u>	**3.5247**	<u>3.2110</u>	3.2715	-
Ours TBEFN	**2.5370**	**2.7680**	<u>3.7442</u>	**2.9368**	**3.0447**	2.2623



(a) input (b) LIME (c) Ying (d) BIMEF (e) Cai

Fig. 9. An image from FUSION dataset. The proposed method has the best transition effect around the edges of high-contrast area, whilst maintaining the details in both background and foreground. More examples will be found in the supplementary material or visiting our github.

(a) (b) (c) (d) (e) (f)

Fig. 11. Ablation experiment on Cai's dataset, where a-f corresponds to sequential number of that in Table III. Our two-branch structure balances the high contrast areas and noisy regions and yield a well-illuminated image with vivid color and fine details. Best zoom in to see more details.

TABLE III
QUANTITATIVE EVLUATION OF ABLATION EXPERIMENTS ON THE STRUCTURE AND LOSS FUNCTION. THE LAST ROW OF THIS TABLE IS PERFORMED ON A GLOBAL END-TO-END FINE-TUNE PROCEDURE. THE OTHER TWO-BRANCH MODELS (D AND E) ARE PERFORMED WITHOUT FINE-TUNING.

	-1E branch only	-2E branch only	Loss function	Cai-2E	Cai-1E	LOL
a	√		l1	18.75/0.733/3.044	18.55/0.784/2.782	16.41/0.456/9.127
b	√		SSIM	19.50/0.787/3.136	19.71/0.851/**2.876**	14.96/0.547/8.705
c		√	SSIM	19.72/0.820/**2.949**	19.08/0.846/2.927	**17.49**/0.720/4.577
d	√	√	SSIM	19.61/0.825/3.010	18.97/0.856/2.994	16.98/0.732/3.229
e	√	√	SSIM+VGG	19.85/0.829/2.967	19.25/0.859/2.981	17.14/0.744/3.232
f(Fine-tune)	√	√	SSIM+VGG	**19.91**/**0.830**/2.970	19.34/**0.863**/2.950	17.31/**0.774**/**3.044**

varied scenes. Figure 9 demonstrates the visual effect of the proposed method. More details will be found in the supplementary material.

V. Discussion

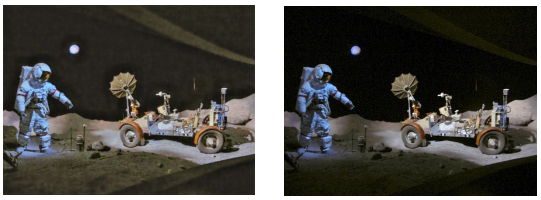
A. Ablation Experiment

Since our model is carried out on a two-branch structure, we performed the ablation experiment on the structure itself and the loss function. The ablation experiments are performed on a mixture of all three datasets. Quantitative and visual effect of the ablation test are shown in Table III and Figure 10 respectively.

Similar to the quantitative evaluation in Table II, the models trained with one branch tend to have bias on a specific illumination level, or would not behave well under all three testing environments. It is easy to show how the judging module works under varied situations. Images enhanced in the -1E branch provides more color information and regional details, while -2E branch are more powerful to dark zones with heavy noise, and provides basic global illumination levels for the whole enhancing process. We marked two patches in Figure 10 for a better comparison. Although any of the single branch is not generalized perfectly, the two-branch structure learns to have the merits of both sides combined. The fusion process is in consistency with the human experience that more trails are more likely to generate a satisfactory result. However, we also noticed that to compensate for the robustness in LOL dataset, where images are all taken in extremely low-light conditions in-doors, performance in the Cai's is sacrificed to some extent.

B. Limitations

.Unlike other end-to-end training strategy, our model need to be trained with a three-stage strategy, meaning that the training process is relatively arduous. However, the training process itself is not time consuming and convergences quickly for each of the stage. Moreover, each stage of the trained parameters is available on our website, which could be used for further study. We also found that this model is slightly vulnerable to dark environment with little variation in neighboring pixels; see Figure 7(3i) and Figure 11, where halos are observed around



(a) (b)

Fig. 10. The proposed method is more vulnerable to dark zone with small variations, which would be over-enhanced and left with halos. (a): Ours, (b) Result of BIMEF[45].

the edges. To handle with this problem is also part of our future work.

VI. Conclusion

In this paper, we proposed a two-stage fusion strategy for low-light images. First, we discussed the relationship of existing enhancing methods and further extended the Retinex theory to a more complicated situation where noise is considered. Then, our enhancing structure is proposed based on the theoretical analysis and motivations. In the proposed structure, images with exposure level of -1E and -2E are first enhanced separately in their own branches to provide possible trails for the judging module, who is dedicated to have them fused with the recognized patterns from the input. Besides, we discussed and verified the introduction of SSIM loss and VGG loss in our experiment, which could be also used in other tasks. Our two-stage denoising and fusion strategy are shown to be both effective and robust. Finally, ablation experiments and limitations are provided for future study.

References

[1] Kinoshita, Y. and H. Kiya, *Scene Segmentation-Based Luminance Adjustment for Multi-Exposure Image Fusion.* IEEE Transactions on Image Processing, 2019.

[2] Prabhakar, K.R., V.S. Srikar, and R.V. Babu. *DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs*. in *ICCV*. 2017.

[3] Pisano, E.D., et al., *Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms.* Journal of Digital imaging, 1998. **11**(4): p. 193.

[4] Celik, T. and T. Tjahjadi, *Contextual and variational contrast*

enhancement. IEEE Transactions on Image Processing, 2011. **20**(12): p. 3431-3441.

[5] Lee, C., C. Lee, and C.-S. Kim, *Contrast enhancement based on layered difference representation of 2D histograms.* IEEE transactions on image processing, 2013. **22**(12): p. 5372-5384.

[6] Huang, S.-C., F.-C. Cheng, and Y.-S. Chiu, *Efficient contrast enhancement using adaptive gamma correction with weighting distribution.* IEEE transactions on image processing, 2012. **22**(3): p. 1032-1041.

[7] Land, E.H., *The retinex theory of color vision.* Scientific american, 1977. **237**(6): p. 108-129.

[8] Jobson, D.J., Z.-u. Rahman, and G.A. Woodell, *Properties and performance of a center/surround retinex.* IEEE transactions on image processing, 1997. **6**(3): p. 451-462.

[9] Jobson, D.J., Z.-u. Rahman, and G.A. Woodell, *A multiscale retinex for bridging the gap between color images and the human observation of scenes.* IEEE Transactions on Image processing, 1997. **6**(7): p. 965-976.

[10] Ying, Z., et al. *A new low-light image enhancement algorithm using camera response model*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[11] Dong, X., Y.A. Pang, and J.G. Wen. *Fast efficient algorithm for enhancement of low lighting video*. in *ACM SIGGRAPH 2010 Posters*. 2010. ACM.

[12] Dong, X., et al. *Fast efficient algorithm for enhancement of low lighting video*. in *2011 IEEE International Conference on Multimedia and Expo*. 2011. IEEE.

[13] Wang, A.Y.-f., B.H.-m. Liu, and C.Z.-w. Fu, *Low-light Image Enhancement via the Absorption-Light-Scattering-Model.* IEEE Transactions on Image Processing, 2019.

[14] Cai, J., S. Gu, and L. Zhang, *Learning a deep single image contrast enhancer from multi-exposure images.* IEEE Transactions on Image Processing, 2018. **27**(4): p. 2049-2062.

[15] Wang, W., et al. *GLADNet: Low-light enhancement network with global awareness*. in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 2018. IEEE.

[16] Wei, C., et al., *Deep retinex decomposition for low-light enhancement.* arXiv preprint arXiv:1808.04560, 2018.

[17] Jiang, Y., et al., *EnlightenGAN: Deep Light Enhancement without Paired Supervision.* arXiv preprint arXiv:1906.06972, 2019.

[18] Wang, R., et al. *Underexposed Photo Enhancement Using Deep Illumination Estimation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

[19] He, R., et al. *Single image dehazing with white balance correction and image decomposition*. in *2012 International Conference on Digital Image Computing Techniques and Applications (DICTA)*. 2012. IEEE.

[20] Ren, J., J. Liu, and Z. Guo, *Context-aware sparse decomposition for image denoising and super-resolution.* IEEE Transactions on Image Processing, 2012. **22**(4): p. 1456-1469.

[21] Jia, X., et al., *An extended variational image decomposition model for color image enhancement.* Neurocomputing, 2018. **322**: p. 216-228.

[22] Guo, X., Y. Li, and H. Ling, *LIME: Low-light image enhancement via illumination map estimation.* IEEE Transactions on image processing, 2016. **26**(2): p. 982-993.

[23] Fu, X., et al., *A fusion-based enhancing method for weakly illuminated images.* Signal Processing, 2016. **129**: p. 82-96.

[24] Fu, X., et al. *A weighted variational model for simultaneous reflectance and illumination estimation*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[25] Ren, W., et al., *Low-Light Image Enhancement via a Deep Hybrid Network.* IEEE Transactions on Image Processing, 2019. **28**(9): p. 4364-4375.

[26] Wang, S., et al., *Naturalness preserved enhancement algorithm for non-uniform illumination images.* IEEE Transactions on Image Processing, 2013. **22**(9): p. 3538-3548.

[27] Shen, L., et al., *Msr-net: Low-light image enhancement using deep convolutional network.* arXiv preprint arXiv:1711.02488, 2017.

[28] Zhang, Y., J. Zhang, and X. Guo, *Kindling the Darkness: A Practical Low-light Image Enhancer.* arXiv preprint arXiv:1905.04161, 2019.

[29] Chen, C., et al. *Learning to see in the dark*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[30] Lore, K.G., A. Akintayo, and S. Sarkar, *LLNet: A deep autoencoder approach to natural low-light image enhancement.* Pattern Recognition, 2017. **61**: p. 650-662.

[31] Lv, F., et al. *MBLLEN: Low-Light Image/Video Enhancement Using CNNs*. in *BMVC*. 2018.

[32] Ignatov, A., et al. *DSLR-quality photos on mobile devices with deep convolutional networks*. in *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[33] Lee, S., G.H. An, and S.-J. Kang, *Deep chain HDRI: Reconstructing a high dynamic range image from a single low dynamic range image.* IEEE Access, 2018. **6**: p. 49913-49924.

[34] Debevec, P.E. and J. Malik. *Recovering high dynamic range radiance maps from photographs*. in *ACM SIGGRAPH 2008 classes*. 2008. ACM.

[35] Ignatov, A., et al. *WESPE: weakly supervised photo enhancer for digital cameras*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018.

[36] Ronneberger, O., P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.

[37] Dabov, K., et al., *Image denoising by sparse 3-D transform-domain collaborative filtering.* IEEE Transactions on image processing, 2007. **16**(8): p. 2080-2095.

[38] Chatterjee, P., et al. *Noise suppression in low-light images through joint denoising and demosaicing*. in *CVPR 2011*. 2011. IEEE.

[39] Qiao, K., et al. *Research on noise testing and reduction of low illumination imaging module*. in *Fifth Symposium on Novel Optoelectronic Detection Technology and Application*. 2019. International Society for Optics and Photonics.

[40] Yan, Q., et al. *Multi-Scale Dense Networks for Deep High Dynamic Range Imaging*. in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2019. IEEE.

[41] Zhang, Y., et al. *Residual dense network for image super-resolution*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[42] Wang, Z. and Q. Li, *Information content weighting for perceptual image quality assessment.* IEEE Transactions on Image Processing, 2010. **20**(5): p. 1185-1198.

[43] Kingma, D.P. and J. Ba, *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

[44] Glorot, X. and Y. Bengio. *Understanding the difficulty of training deep feedforward neural networks*. in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010.

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

[46] Li, M., et al., *Structure-revealing low-light image enhancement via robust Retinex model.* IEEE Transactions on Image Processing, 2018. **27**(6): p. 2828-2841.

[47] Lee, C., C. Lee, and C.-S. Kim. *Contrast enhancement based on layered difference representation*. in *2012 19th IEEE International Conference on Image Processing*. 2012. IEEE.

[48] Ma, K., K. Zeng, and Z. Wang, *Perceptual quality assessment for multi-exposure image fusion.* IEEE Transactions on Image Processing, 2015. **24**(11): p. 3345-3356.

[49] Mittal, A., R. Soundararajan, and A.C. Bovik, *Making a "completely blind" image quality analyzer.* IEEE Signal Processing Letters, 2012. **20**(3): p. 209-212.