Learning an Adaptive Model for Extreme Low-light Raw Image Processing

Qingxu Fu¹, Xiaoguang Di²*, Yu Zhang²

¹Control and Simulation Center, Harbin Institute of Technology, Harbin, 150080, People’s Republic of China
²National Key Laboratory of Tunable Laser Technology, Harbin Institute of Technology, Harbin, 150080, People’s Republic of China
*E-mail: dixiaoguang@hit.edu.cn

Abstract: Low-light images suffer from severe noise and low illumination. Current deep learning models that are trained with real-world images have excellent noise reduction, but a ratio parameter must be chosen manually to complete the enhancement pipeline. In this work, we propose an adaptive low-light raw image enhancement network to avoid parameter-handcrafting and to improve image quality. The proposed method can be divided into two sub-models: Brightness Prediction (BP) and Exposure Shifting (ES). The former is designed to control the brightness of the resulting image by estimating a guideline exposure time $t_1$. The latter learns to approximate an exposure-shifting operator $ES$, converting a low-light image with real exposure time $t_0$ to a noise-free image with guideline exposure time $t_1$. Additionally, structural similarity (SSIM) loss and Image Enhancement Vector (IEV) are introduced to promote image quality, and a new Campus Image Dataset (CID) is proposed to overcome the limitations of the existing datasets and to supervise the training of the proposed model. In quantitative tests, it is shown that the proposed method has the lowest Noise Level Estimation (NLE) score compared with BM3D-based low-light algorithms, suggesting a superior denoising performance. Furthermore, those tests illustrate that the proposed method is able to adaptively control the global image brightness according to the content of the image scene. Lastly, the potential application in video processing is briefly discussed.

1 Introduction

Images play an irreplaceable role in industry, military and entertainment. As the art of light, how to obtain better images in the low-light environment has been extensively studied in the literature. Some advanced photography equipment can help but at an expensive cost. On the other hand, there are few limitations to post-process low-light images. However, it is challenging due to problems such as noise, color distortion et al.

Exposure time, ISO (which measures the sensitivity of the image sensor) and aperture are known as three pillars of photography. With extended exposure time, the low-light problem can be easily solved, but it is not realistic because if the camera is not fixed or the scene contains moving objects. Higher ISO introduces more noise and is not always available for mobile devices. As a flexible solution, low-light image enhancement provides an alternative for imaging in the low-light environment.

In general, current low-light image enhancement method can be classified as follows:

Classic low-light image enhancement methods, with no application of convolutional neural networks. Histogram equalization (HE)[1] and gamma correction[2] provide simple solution for low-light image processing. Based on Retinex theory[3], Single-scale Retinex (SSR)[4], Multi-scale Retinex (MSR)[5], SRIE[6] and LIME[7] were proposed, enhancing low-light image by estimating the illumination map and reflectance map. Dehazing[8] methods can also be applied to this subject, regarding low-light images as inverted hazed image.

With the rapid development of the deep neural network, researchers have combined classic low-light image enhancement methods with Convolutional Neural Networks (CNNs). Chen et al. proposed Retinex-Net[9] to decompose low-light image into illumination and reflectance and use BM3D[10] for denoising.

Those methods assume low-light images are free of noise or noise is already removed by additional denoising process. Therefore, when applying those methods on real-world low-light images, an additional denoising process is required. However, most low-light images suffer from severe noise, traditional denoising methods like BM3D are not able to provide satisfactory results. To be rid of image noise, researchers also explored the method of deep learning[11–14] in denoising topics.

There are some low-light images which are barely visible before post-processing, as they are captured in extremely dark environments or with very short exposure time. In this work, those images are named as extreme low-light images. While the classic methods are unable to tackle the severe noise and serious color distortion in those images, it is then discovered that with the help of deep learning and raw-format image datasets, we can still obtain satisfactory results. Chen et al. proposed an end-to-end solution for raw low-light image enhancement[15], dealing with the low-light problem and denoising in a single model. However, extra brightness amplification ratio is needed and requires manual parameter tuning for different scenes, gaining excellent denoising effects but losing the flexibility and adaptability of classic methods.

Currently, all datasets available for low-light enhancement models chose a longer-exposed image as the ground-truth image in training, which was heavily dependent on the dataset collector’s experience to select exposure parameters such as exposure time, ISO and white balance mode. Thus these parameters were usually not optimal, considering the content of the scene, illumination of the environment et al., which can lead to an undesired trained model.

It is hard for deep CNNs to learn how we determine the best exposure parameters for the ground-truth images because this can be very subjective and lacks specific standards. It can lead to more
problems if there is more than one dataset collector. On the other hand, making a baseline for ground-truth selection is equally difficult. Nevertheless, it is possible to predict how an image varies when exposure parameters change. We name this process Exposure Shifting (ESN). With a learned ES model, it becomes possible to reverse study what good exposure parameters (e.g. exposure time) should be for each low-light image, utilizing the characteristics of the back-propagation algorithm.

The contributions of our work are summarized as follows:
1) We presented Campus Image Dataset (CID), which contains a variety of scenes captured by multi-level exposure. CID provides powerful support for the training of our data-driven low-light enhancement model.
2) We proposed an adaptive two-stage low-light image enhancement model, which provides state-of-the-art low-light noise suppression as well as adaptive global brightness control effect.

In stage one, a Brightness Prediction Network (BPN) was introduced to estimate a proper exposure time based on the content of images as well as Exif (Exchangeable image file) metadata. BPN was trained to provide adaptive brightness control during image enhancement and to get rid of the external parameters required in the Exposure Shifting stage.

In stage two, an Exposure Shifting Network (ESN) was proposed to estimate a longer-exposed version of the image with target exposure time estimated by BPN. Moreover, ESN also completes image denoising and ends up with the final resulting RGB image.

The rest of the article is organized as follows. In section Related Works, previous research works are provided. The Dataset section demonstrates the importance and advantages of our CID (Campus Image Dataset) dataset. In the Method section, our model and its training details are provided, as well as the design concept of the model. The proposed model is compared with other state-of-the-art methods in Experiments section, from multiple perspectives including denoising, adaptive brightness control performance and computational cost. The possible extension in video processing is briefly discussed here. In the Conclusion and Discussion section, the advantages and the limitations of our model are concluded. We also present the future expectations of this work.

2 Related Works

The method we proposed was inspired not only but the application of deep neural networks, but also by classic methods based on the Histogram Equalization model, dehazing model and Retinex model.

2.1 Histogram Equalization Methods

Histogram Equalization (HE) method is used extensively to process digital images. The basic idea is to improve image visibility by changing the image histogram into a uniform distribution, which effectively increases image entropy for low-light images. In [16] it is argued that a color-invariant representation can be obtained by applying HE. However, HE tends to cause noise amplification, details disappearance and color distortion in low-light image enhancement.

On account of the drawbacks of HE, a number of improvements have been put forward. Adaptive Histogram Equalization (AHE) [17] and its variation Contrast-limited Adaptive Histogram Equalization (CLAHE) [18] perform the histogram equalization in a region surrounding each pixel, instead of across the whole image. Hue-preserving color image enhancement [19] works on contrast-enhancing and hue-preserving. [20] and [21] on preserving the original image luminance.

2.2 Dehazing and Retinex Model-Based Methods

A number of low-light enhancement methods are based on the dehazing and Retinex model. It is observed that low-light images and inverted haze images share many similarities. For this reason, [8] presented a method in which low-light image is inverted, dehazed and then inverted back again. This method was studied further by [22] and [23].

On the other hand, the Retinex model [3] revealed that a natural image is made up of the illumination map and the reflectance map. With Retinex decomposition and the assumption that the illumination map is smooth, researchers have proposed single-scale Retinex [24] and multi-scale Retinex [4]. Based on image fusion, adjustments can be applied to the illumination map to improve the performance [25]. [26] focused on preserving natural characteristics with a Bright-Pass filter. In [27], a variational Retinex model was proposed and the illumination estimation problem was formulated as a Quadratic Programming optimization problem. In [28] illumination map is estimated considering local structure and [29] focused...
on revealing the structure details from the reflectance map. A recent study shows that the Retinex model can combine with the atmospheric scattering model[30]. More refined models were presented by [31, 32] based on variational Retinex theory.

However, noise modeling is not well established in those algorithms. In [8] it is assumed that noise is insignificant and can be removed before or after the Dehazing stage. And the Retinex based models rely on classic noise reduction algorithms (e.g. BM3D) to reduce noise found in the reflectance map. Therefore it can achieve good results in mild low-light images but tends to amplify noise when processing severe or extreme low-light images. Besides, the variational Retinex model is very time-consuming as it takes many iterations to solve the optimization objectives and so are many denoising algorithms, thus it is hard to achieve real-time processing.

2.3 Data-driven Methods

Image denoising and low-light enhancement can be realized with the data-driven approach, separately or integrated.

Noise reduction has been the subject of extensive studies, [11] has discovered that convolutional neural networks can compete with the state-of-the-art classic denoising algorithm BM3D. The works on date-driven denoising have been extended by [14, 33, 34]. Although these works were not targeted at low-light image processing, they provide some references for the research of other image processing studies.

The first application of the data-driven method is LLNET [12], which utilizes a deep auto-encoder to learn from synthetic low-light image datasets and the object of denoising is integrated into the low-light enhancement process. Also, it was demonstrated that Retinex based methods can be further improved by deep learning [9, 35], but these works still have undesirable denoising performance because synthetic datasets are used and they cannot reflect the characteristics of real low-light images.

To solve the drawbacks of synthetic datasets, a large multi-exposure image dataset was proposed in [36], but it was found by [15] that models that utilize raw-format images can provide much better results in low-light enhancement. In [15], it suggested that the state-of-the-art results can be achieved by using raw-format paired images and training in an end-to-end way. However, since the paired two images have different exposure time, a ratio was introduced as an additional input to bridge the gap. Thus after the model is trained and with no reference to indicate the ratio, [15] requires manual adjustment for every different image to be enhanced, resulting in many drawbacks in practical applications.

3 Dataset

For extreme low-light image enhancement, there are few optional datasets. The Google HDR+ dataset[37] and Darmstadt Noise Dataset[38] avoid the disadvantages of synthetic datasets, but they were mainly collected in environments with sufficient illumination. The RENOIR[39] and learning-to-See-In-the-Dark dataset (SID)[15] provides real low-light image pairs, which is captured by carefully selecting the ISO and exposure time for each scene. The Exclusively Dark Dataset (EDD) [40] is another low-light dataset with object-level annotations, which is targeting at object recognition. But these datasets cannot satisfy the need of our study. In our work, in order to adaptively enhance images with different noise levels, it is needed to consider how the exposure parameters have an impact on low-light images. More specifically, as the environment illumination grows lower, low-light images are gradually overwhelmed by noise and invalid pixels. Although it is not flexible to manipulate the environment illumination, the camera exposure time setting can be easily changed to simulate this illumination shift.

Therefore, it is expected that there is a dataset containing, instead of just some image pairs, a number of multi-exposed image series, in each of which there are low-light images of different levels and one reference image captured in the same scene.

Overall, there are three conditions need to be met for a satisfying dataset:

- Real scenes. Unlike synthetic images, photos captured in real scenes reflect much more diverse noise and distortion pattern
- Multiple exposure levels. Using a number of multi-exposed image series to train an adaptive model capable of enhancing low-light images of different levels.
- Raw-format images. Most data captured by the camera sensor is lost in format conversion, those lost information is essential for extreme low-light image enhancement.

![CID Image Group](image)

**Fig. 3.** Demonstration of a CID image group.

Based on the above description, we proposed our Campus Image Dataset (CID). There are about 200 image groups in CID, each of which consists of 8 raw images with different exposure time, shot continuously in the same scene using a tripod (Fig.3). To be specific, for every group, an upper limit of exposure time was selected to make sure that the first image in the sequence was slightly overexposed, and then a lower limit was chosen to capture an extreme low-light image as the last image in the sequence, and the gap between the limits was filled up by another 6 images. Finally, the ground-truth image for each group was manually selected, and invalid images that had nothing but noise due to extreme short exposure were tagged and excluded. A demonstration of scenes in CID is presented in Fig.1.

4 Method

4.1 Method Overview

Most of the existing methods are unable to suppress the severe noise present in extreme low-light images. On the other hand, recent studies e.g. [15] have concentrated on the extreme low-light enhancement with raw-format images and real-scene datasets in an end-to-end way, but unlike the classic methods, the brightness of the enhanced image cannot be automatically controlled in this framework, more specifically, an external ratio has to be manually adjusted when the input image has no paired image as the reference.

It is our aim to find a solution to enhance extreme low-light images, combining the advantages of adaptive brightness control and superior denoising performance. It is found that the problem of low-light enhancement can be considered as a special case of changing the time of exposure. With $t_0$ and $t_1$ representing exposure time, the low-light image $X_{t_0}$ can be interpreted as a normal image $Y_{t_1}$ corrupted by a hypothetical exposure-shifting operator $ES$, which only alters the exposure time but keeps the aperture and ISO unchanged:

$$ BX_{t_0} = ES(Y_{t_1}, t_0) $$

(1)

In this process, more stochastic noise $N$ is generated by operator $ES$ because $t_0 < t_1$, corresponding to the low PSNR (Peak Signal to Noise Ratio) value in low-light images. Contrariwise, the corrupted
low-light image can also be restored by the same operator $ES$:

$$\hat{Y}_{t_1} = ES(X_{t_0}, t_1)$$

(2)

The noise $N$ is reduced because $t_1 > t_0$. Thus the normal image $Y_{t_1}$ is reconstructed.

The latter process, named as Exposure Shifting (ES), can enhance low-light images as well as suppress image noise. More importantly, it can be learned in a data-driven and end-to-end way, thus it is expected to have superior denoising performance. However, a proper guideline exposure time $t_1$ always has to be provided in the reconstruction $Y_{t_1}$. To make our model adaptive, a Brightness Prediction (BP) procedure is introduced to achieve guideline time estimation.

In brief, the extreme low-light enhancement process is split into two procedures. Firstly, for a raw-format low-light image (with real exposure time $t_0$), an optimal guideline exposure time $t_1$ is estimated via a sub-model called Brightness Prediction Network (BPN). Secondly, the final RGB-format enhanced image is obtained by another sub-model, namely Exposure Shifting Network (ESN), using estimated guideline time $t_1$ to control result brightness.

In addition, besides the image itself, some extra data are involved in both BPN and ESN, such as white balance index, ISO, $t_0$, and $t_1$. The Image Enhancing Vector (IEV) is put forward to encode the involved extra data.

### 4.2 Image Enhancing Vector

Raw image array is the original data captured by the camera. Other key parameters, such as camera white balance, ISO, exposure time, time and date, are stored together with the raw image array as Exif metadata. In researches of HDR imaging[41], exposure time and ISO information have played an essential role in calculating the response curve and LDR to HDR conversion. Although it is not necessary to explicitly train the network to learn the camera response curve, feeding these additional parameters into the network as input avoids underfitting caused by insufficient features.

Image Enhancing Vector (IEV) is proposed to introduce extra features into the convolution neural networks. In this paper, IEV consists of ISO $u_s$, white balance indices for 4 raw image channels ($w_r, w_g, w_b, w_{g2}$), exposure time of the low-light image $t_0$ and guideline exposure time $t_1$. Moreover, it is possible to append other Exif metadata into IEV to improve network performance when necessary, e.g. aperture and focal length.

The IEV is used as network input in both BPN (for $t_1$ estimation) and ESN (for exposure shifting), but a difference exists. As it is shown in Fig.2, in ESN procedure the IEV can be written as:

$$Y^{t_0 \rightarrow t_1} = Y(t_0, t_1) = (w_r, w_g, w_b, w_{g2}, u_s, ..., t_0, t_1)$$

(3)

Where $(w_r, w_g, w_b, w_{g2})$ are the white balance indices of 4 raw image channels and $u_s$ is the ISO value (controlling the camera sensitivity).

In BP procedure the $t_1$ in IEV is removed. It is renamed as partial IEV (pIEV) to indicate the difference:

$$Y^{t_0 \rightarrow t_1}_p = Y_p(t_0) = (w_r, w_g, w_b, w_{g2}, u_s, ..., t_0)$$

(4)

IEV and pIEV are fed into the convolutional neural network as additional channels of the raw image array. For example, the first extra channel provided by IEV is a uniform image filled up with pixel value $u_r$.

### 4.3 Exposure Shifting

Exposure Shifting (ES) is the second procedure in our model but is trained first. The U-NET[42] is selected as the structure of the Exposure Shifting Network (ESN). Compared with the fully convolutional network, the U-NET architecture reduced the usage of GPU memory, thus it can easily process images with high resolution.

Recall that the basic unit of our Campus Image Dataset (CID) is an image group, which consists of 8 raw-format images with different exposure time but captured in the same scene. Let $(X_R, Y_R)$ be the paired raw image extracted from one group, in which $Y_R$ is the pre-selected ground-truth image and $X_R$ is a random-selected low-light image. More specifically, $X_R$ is randomly chosen within this image group, with $Y_R$ and over-exposed images excluded (Fig.3). Moreover, the exposure time of $X_R$ is denoted as $t_0$. The corresponding RGB-format image of $Y_R$ is denoted as $Y$ and its exposure time as $t_R$.

It is expected for the ESN to learn the exposure-shifting operator $ES$ from a large number of paired images. Specifically, the ESN is required to estimate an image $\hat{Y}$ that resembles the $Y$ as closely as possible. In this way ESN successfully changes the raw-format low-light image $X_R$ into a RGB-format longer-exposed version of itself. More importantly, the serious noise of $X_R$ can also be removed. This procedure can be written as:

$$\hat{Y}^{\Theta}_{t_0} = F_{ES}(X_R, V^{t_0 \rightarrow t_R}_{t_0} | \Theta_1)$$

$$V^{t_0 \rightarrow t_R}_{t_0} = V(t_0, t_1 = t_R)$$

(5)

Where ESN is denoted as $F_{ES}$, its parameters as $\Theta_1$. Note that since the BP procedure has not yet involved, the guideline exposure time $t_1$ is replace by $t_R$ in IEV $V^{t_0 \rightarrow t_R}_{t_0}$. The enhanced result image is represented by $\hat{Y}^{\Theta}_{t_0}$.

Two metrics to guide the training process of the ESN model are Mean Square Error (MAE) and multi-scale Structural Similarity (SSIM)[43]. The former is considered as a simple but effective metric while the latter is used to measure the similarity between the model’s output and the ground-truth image.

The final MAE loss is the average of the channel MAE loss, for simplicity, the image channel is not presented in the equation. The calculation of the multi-scale structural similarity (SSIM) can be referred in [43]. With subscript $t_0$ omitted, the loss function is defined as the linear combination of $L_{MAE}$ and $L_{SSIM}$:

$$L_{ES} = L_{MAE} \left( \hat{Y}^{\Theta}_{t}, Y \right) + \alpha L_{SSIM} \left( \hat{Y}^{\Theta}_{t}, Y \right)$$

(7)

Where $\alpha$ is a constant and $0 \leq \alpha < 1$. The influence of $\alpha$ will be discussed in the Experiments section.

The parameters of the ESN network $\Theta_1$ are learned by minimizing $L_{ES}$ over $K$ pairs of images in the training set:

$$\Theta_1^* = \argmin_{\Theta_1} \sum_{k=1}^{K} L_{ES} (\hat{Y}^{\Theta}_{k}, Y_k)$$

(8)
The purpose of BPN is to control the brightness of the final result image, and as it is revealed in the S-shaped curve because the pixel brightness is limited within [0, 1] (or [0, 255] for 8-bit image storage) and the camera sensor is less sensitive to the change of \( t \) when the pixel value is close to 0 or 1. In other words, as the exposure time \( t \) setting changes, some pixels are not sensitive and there are little changes in the pixel brightness compared with some other pixels. For example, when photographing a street at night, pixels revealing the lamp bulbs are always saturated unless the \( t \) is extremely small, and pixels representing the dark sky are always close to zero unless it is affected by noise. If those pixels are involved in BPN training, they can cause gradient vanishing in the back propagation algorithm because \( \frac{\partial B_{Br}(\epsilon)}{\partial \epsilon} = \frac{\partial B_{Br}(\epsilon)}{\partial t} \cdot R \) and as it is revealed in the S-shaped curve.
softmax function: Combining the Equation 12 and 13, they can be simplified by the softmax function:

$$W = \text{softmax} \left( -\frac{(Y - \mu_w)^2}{2\sigma_w^2} \right)$$

Lastly, the loss function $L_{BP}$ is designed to check if the final result image $\hat{Y}^{\Theta_2;\Theta_1^*}$ has the same AoI. Having the same AoI means that in this AoI area $\hat{Y}^{\Theta_2;\Theta_1^*}$ and $\mathcal{Y}$ share similar brightness. Moreover, recall that pixels outside AoI are not sensitive to the change of exposure time $t$ or the ES process, thus it can be derived that the overall image brightness of $\hat{Y}^{\Theta_2;\Theta_1^*}$ resembles that of $\mathcal{Y}$. In this way, BPN completed the adaptive Brightness Prediction objective.

Then enhanced image $\hat{Y}^{\Theta_2;\Theta_1^*}$ is converted into a grayscale image $\hat{Y}^{\Theta_2;\Theta_1^*}$. When the AoI of $\hat{Y}^{\Theta_2;\Theta_1^*}$ and $\mathcal{Y}$ overlaps exactly, the following loss function reaches its minimum:

$$L_{BP} = L_{BP} \left( \hat{Y}^{\Theta_2;\Theta_1^*}, \mathcal{Y} \middle| \Theta_2; \Theta_1^* \right)$$

$$= -\frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \exp \left( -\frac{(Y_{G,i,j} - \mu_w)^2}{2\sigma_w^2} \right) W_{ij}$$

Where $\sigma_w^2 = 0.04$ is another variance constant. The image value at position $(i, j)$ is denoted by $Y_{G,i,j}$ and $W_{ij}$ overlaps exactly. The following loss function reaches its minimum:

$$\hat{Y}^{\Theta_2;\Theta_1^*} = \arg\min_{\hat{Y}^{\Theta_2;\Theta_1^*}} \sum_{k=1}^{K} L_{BP} \left( \hat{Y}^{\Theta_2;\Theta_1^*}, W_k \middle| \Theta_2; \Theta_1^* \right)$$
4.5 Method Summary and Discussion

In Algorithm 1, the training process of our model is summarized, together with the method to evaluate and apply this model.

Algorithm 1 Model training and evaluation

1: Initialize ESN, BPN parameters $\Theta_1$ and $\Theta_2$
2: Prepare CID dataset
3: function TRAIN(Dataset)
4: for EpochTrainESN = 1 → $E_1$ do
5: for Image Pair $k = 1 → K$ do
6: Read raw image pairs $(X_{R,k}, Y_{R,k})$ and $Y_{t^{0} → t^{g}}$
7: Convert to RGB format $Y_k^* \leftarrow Y_{R,k}$
8: Compute $\hat{Y}_k^*$ via Equation 5
9: $\Theta_1 \leftarrow \arg \min_{\Theta_1} \sum_{k=1}^{K} L_{ES} \left( \hat{Y}_k^*, Y_k \right)$
10: end for
11: end for
12: $\Theta_1^* \leftarrow \Theta_1$
13: for EpochTrainBPN = 1 → $E_2$ do
14: for Image Pair $k = 1 → K$ do
15: Read raw image pair $(X_{R,k}, Y_{R,k})$ and $Y_{t^{0} → t^{g}}$
16: Compute $\hat{Y}_k^*$ and $\hat{Y}_k^*$ via Equations 9 and 10
17: Convert to grayscale image $Y_{G,k}^r \leftarrow Y_{R,k}$
18: Convert to grayscale image $\hat{Y}_{G,k}^r \leftarrow \hat{Y}_k^*$
19: Compute $W_{k}^r$ via Equation 14
20: $\Theta_2 \leftarrow \arg \min_{\Theta_2} \sum_{k=1}^{K} L_{BP} \left( \hat{Y}_{G,k}^r, W_k \left| \Theta_2; \Theta_1^* \right. \right)$
21: end for
22: end for
23: $\Theta_2^* \leftarrow \Theta_2$
24: return $\Theta_1^*, \Theta_2^*$
25: end function
26: 
27: function EVALUATE(Image)
28: Read raw-format image $X_{R}$
29: Read all elements in $Y_{t^{0}}$ from raw-image file
30: BPN: Compute $\hat{Y}_k^{t^{2}} \leftarrow F_{BP} \left( X_{R}, Y_{t^{0}} \left| \Theta_2^* \right. \right)$
31: ESN: Compute $\hat{Y}_k^{t^{2}} \leftarrow F_{ES} \left( X_{R}, Y_{t^{0} → t^{g}} \left| \Theta_1^* \right. \right)$
32: return $\hat{Y}_k^{t^{2}}$
33: end function

As mentioned earlier in the last section, training BPN with input $X_{R}$ and label $t_{g}$ is straightforward but not feasible. First let $F_{BP}$ be the hypothetical optimal BPN model, then $\hat{Y}_k^{t^{2}}(X_{R}) = F_{BP}(X_{R})$ is the optimal estimation of guideline exposure time. For a pair of image and time label $(X_{R}, t_{g})$, the $t_{g}$ can be considered as one sampling from a Gaussian distribution: $t_{g} \sim N \left( \hat{t}_1, \sigma_1^2 \right)$. The variance $\sigma_1^2$ in this process is too large because:

- In every CID image group, the ground-truth image is chosen manually and is dominated by subjective human judgments.
- The ground-truth image is selected from only 8 candidate images, but the optimal $\hat{t}_1$ exists somewhere in the time continuum $(0, \inf)$.

Furthermore, the number of samples is very limited since each image group can only provide one $t_{g}$ sample. As a result, this straightforward approach cannot be applied in the training of BPN due to the high variance of the label $t_{g}$ and the limited number of samples.

5 Experiments

5.1 Training

Limited by the GPU memory, we did not adopt the Batch Normalization [45] technique. To accelerate model training, before the training process started, each channel of the input image, as well as each element of IEV, was normalized along the sample dimension. Then the means and variances used in the normalization became invariant constants. Additionally, images for training were randomly cropped into $512 \times 512$ patches.

Both ESN and BPN were trained with Adam optimizer[46], following the procedures shown in Algorithm 1. The ESN was trained first with the learning rate slowly descending from $2 \times 10^{-4}$ to $1 \times 10^{-5}$. After about 300 epochs, when the loss on training sets became stable. Then the parameters in ESN were made untrainable. Then BPN was appended to the training graph and trained with the same learning rate. Considering BPN may require global information to decide the proper exposure time, we used bigger patches but avoid training with the whole image to prevent overfitting. Although the loss must propagate through the deep ESN before reaching BPN, the training process went smoothly and completed after 100 epochs.

5.2 Content-Based Brightness Control

Firstly, we highlight the adaptive brightness control feature of our method. Here we explicitly define the brightness as the mean value of an image $\hat{Y}$, denoted as $B_r(\hat{Y})$. We expect a dynamic adjustment of the image brightness based on the image content. More specifically, while extreme low-light images suffer from severe noise along with color distortion and are barely visible before post-processing, other mild low-light images only have moderate noise and relatively lower brightness compared with normal images. Thus, the enhancement algorithm needs to make adjustments adaptively for different input images. Currently the classic methods, such as algorithms based on Retinex and Dehazing, have poor performance on the extreme low-light images, on the other hand, learning-based methods targeting at end-to-end denoising can process all kinds of low-light images, but they lack the flexibility and adaptability because they either need brightness amplification hand-tuning, or simply serve as a denoising procedure in other low-light methods. The proposed model is of the advantages of both and is able to process low-light images of different levels.

In order to evaluate the adaptability, we apply the trained model to enhance all images in the test set. Recall that the image group is the basic unit of our CID dataset, accordingly, instead of evaluating single images, the collaborative characteristics across different images are investigated within each image group. Note that the group identity of an image is NOT involved in model training and only serves as a tag to gather resulting images for further analysis.

For a sequence of multi-exposed images $(X_{R,1}, X_{R,2}, \ldots)$ with increasing exposure time, it is expected that after enhancement the resulting images possess similar image brightness, namely $B_r(\hat{Y}_1) \approx B_r(\hat{Y}_2) \approx \ldots$, where $\hat{Y}$ denotes the enhanced image. Fig.5(a) illustrates a busy-road scene, the first row displays the RGB images produced by the camera, the second displays the enhanced results by utilizing our method on individual raw images. Note that the first image in this group is exposed for 1/20 seconds, causing motion blur to the truck, while the last image is exposed for only 1/100 seconds. It is shown that all four images have approximately uniform brightness after enhancement, despite the changing $t_{g}$ and the moving vehicle headlight. Fig.5(b) presents the performance of our method on extreme low-light conditions. The last image in the sequence suffers from both low environment illumination and short

<table>
<thead>
<tr>
<th>Methods</th>
<th>SNLE</th>
<th>NLE</th>
<th>NV</th>
<th>Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>4.0689</td>
</tr>
<tr>
<td>MSR+BMDSD</td>
<td>0.0014</td>
<td>1.3792</td>
<td>0.8998</td>
<td>7.0068</td>
</tr>
<tr>
<td>LIME+BMDSD</td>
<td>0.0086</td>
<td>4.2437</td>
<td>10.1789</td>
<td>7.1084</td>
</tr>
<tr>
<td>R-Net+BMDSD</td>
<td>0.0017</td>
<td>0.2467</td>
<td>0.4491</td>
<td>6.1503</td>
</tr>
<tr>
<td>Ours</td>
<td>0.0015</td>
<td>0.1061</td>
<td>0.4554</td>
<td>6.6112</td>
</tr>
</tbody>
</table>
exposure and the color and shape of objects in the image are drowned out by noise. However, the resulting image has the noise reduced and still maintain acceptable global brightness.

As mentioned above, for quantitative analysis, we gathered images that belong to the same groups and calculated the brightness of all images in order to observe the shift of brightness. Let $(Br_{1}, Br_{2}, \ldots)_{k}$ be the image brightness of each unenhanced low-light image in $k$-th image group and let $(Br'_{1}, Br'_{2}, \ldots)_{k}$ be that of the enhanced ones. Then, the mean value and the CV (coefficient of variation) of $(Br_{1}, Br_{2}, \ldots)$ is computed and denoted as $(\mu, cv)_{k}$, correspondingly, $(\mu', cv'_{k})$ is calculated from $(Br'_{1}, Br'_{2}, \ldots)_{k}$. In Fig.8, we plotted $(\mu, cv)$ and $(\mu', cv')$ for all groups. To illustrate the change after enhancement, for each group $(\mu, cv)$ (marked with star) and $(\mu', cv')$ (marked with circle) are linked with a dotted line. It can be observed from Fig.8 that for different groups that captured from diverse scenes, the enhanced images all end up with increased image brightness and significantly reduced CV. And yet the brightness growth between groups can be very different, ranging from around 60% to over 1,000%. Considering together with Fig.exp, Fig.8 indicates that the BP sub-model is able to adaptively control the global image brightness according to not only the brightness of the original image, but also the content of the scene. On the other hand, the dramatic reduction of CV suggests that the enhanced images within each group share almost identical brightness in spite of the varying exposure time $t_0$. To be precise, our model can control group brightness CV under 0.29 on 76.3% test set groups and 98.5% research.
training set groups considering the influence of the diverse scenes, illumination conditions, capture ISO, etc.

When designing the model, we expect the BPN to extract features of the scene content and then estimates a reasonable $t_1$ accordingly. In order to see through clearly how this black box works, Fig.9 is plotted to illustrate the relationship between the real exposure time $t_0$, the BPN-estimated guideline exposure time $t_1$ and the brightness of enhanced images $B(R)$. Similar to Fig.8, we use group average to simplify the figure, in which each point corresponds to an image group. As shown in Fig.9, the network tends to estimate $t_1$ positively associated with $t_0$. Furthermore, for groups with identical average $t_0$, the content of the scenes show an impact on $t_1$ prediction. To be specific, if the scene of a group has more relatively bright areas, a smaller $t_1$ is more likely to be adopted to prevent overexposure and vice versa. And yet a few exceptions exist, suggesting the BPN sub-model has learned some more sophisticated rules in the neural network black box.

### 5.3 Denoising Evaluation

We compare our method with four classic and state-of-art methods, including multiscale Retinex (MSR)[5], illumination map estimation based (LIME)[7], deep Retinex decomposition (Retinex-Net, R-Net)[9] and learning-to-see-in-the-dark (SID)[15]. For fair comparison, 3D transform-domain filtering (BM3D)[10] is applied on MSR, LIME, R-Net, since those methods do not model noise and require extra denoising process. Additionally, we train the SID model on the CID dataset and manually select ratio $\gamma$ for the SID model since it does not provide automatic brightness estimation.

Recall that the BPN utilizes a reduced reference evaluation index as the loss in its training. In this way, the exposure parameters of the ground-truth image, which is chosen by highly subjective human judgment, will have much less influence on BPN model training. As a result of this novel technique, however, PSNR and SSIM evaluation judgment, will have much less influence on BPN model training. As in Fig.6, the enhancement process of each frame is completely independent of other frames, neither the estimated $t_1$ nor the resulting images show any sudden fluctuation. However, for practical considerations, it is also convenient to implement guideline time filters between ESN and BPN, which allows the value of $t_1$ to be dynamically filtered or amplified according to different demands.

### 5.4 Low-light Video Processing

The advantage of our work lays in getting rid of the brightness control ratio while achieving state-of-the-art low-light noise suppression. Without any handcrafting parameters as network input, we can directly apply our method to raw low-light video processing.

A raw-format video can occupy a very large amount of storage space. Limited by devices, we only developed a relatively simple way to verify our method on raw low-light videos. In Fig.10, we experiment on 21 continuous raw images captured in a burst session, shot with short exposure time and identical ISO, to discuss the potential applications in raw video processing. In this scene, the camera starts from a bridge with lighting decoration, then turns to the dark riverbank, and finally to the dimly illuminated sidewalk. As is shown in Fig.10, though the enhancement process of each frame is completely independent of other frames, neither the estimated $t_1$ nor the resulting images show any sudden fluctuation. However, for practical considerations, it is also convenient to implement guideline time filters between ESN and BPN, which allows the value of $t_1$ to be dynamically filtered or amplified according to different demands.

### 5.5 Computational Cost

The proposed algorithm is evaluated on the ModelArts cloud service, which is equipped with 8 vCPUs and an Nvidia-P100 GPU (16GB GPU memory). The execution time for a high-resolution image (3968 $\times$ 2976) is 0.27 seconds on average, which only shows a minor increase compared with the SID algorithm (0.21 seconds). More specifically, the execution time of BPN and ESN sub-model is 0.05 and 0.22 respectively. For comparison, it takes 4.32 seconds for BM3D (GPU implementation [50]) to complete the denoising process, therefore other low-light methods that depend on BM3D for denoising, including MSR, LIME, R-Net et al., are much less efficient on devices with GPUs.

### 6 Conclusion and Discussion

#### 6.1 Conclusion

In this paper, an adaptive raw image enhancement model is proposed for extreme low-light image processing. The two-stage framework provides adaptability to images with different scenes and exposure parameters. We also presented the CID dataset for model training and evaluation. The experimental test shows that our model can provide the state-of-the-art low-light image denoising as well as adaptive global brightness control.
The proposed method has many advantages over existing approaches. In the first place, no external parameters are needed after model training, and the algorithm is free of Exif metadata (e.g. white balance, ISO and exposure time) sufficient to complete the process. Therefore, our model is able to process a large batch of raw images without parameter-handcrafting. Secondly, our model significantly outperforms other methods in which denoising works as a separable step; because instead of simply implementing noise-suppression algorithms (e.g. BM3D) before or after the main low-light enhancement procedure, we make denoising procedure an integrated part in our model, allowing the ESN module to learn exposure shifting and denoising simultaneously in an end-to-end way. In quantitative experiments, we adopted no-reference image quality metrics including SNLE, NLE, NV and image entropy to test the denoising performance. Our model has the smallest noise variance on NLE metric compared to BM3D-based low-light methods and obtains approximate optimal scores on the other three indices. We also reveal the potential application in low-light video processing. The BPN module makes it possible for our model to process images with diverse exposure levels, from extreme low-light images to mild ones. But to be applied in video processing, it should face more challenges, e.g. the brightness of frames should change continuously and avoid fluctuation. We experiment on continuous raw image series, despite that no information of adjacent frames is used, the output of BPN changes continuously with the illumination condition and shows no fluctuation.

6.2 Limitations and Future Works

The model proposed is limited by dataset, which is a common problem in data-driven methods. On the one hand, the model is trained by raw-format and multi-exposed images, which significantly improved image quality. But on the other hand, all reference ground-truth images are captured in the real scenes, as a result, they sometimes have minor defects such as noise, halos around the light source, local over-exposure and white balance deviation. The denoising performance may be further elevated if those drawbacks can be avoided. Then the image groups in CID have not covered enough nature scenes. For example, it is found that green grass and trees tend to have low color saturation in the resulting images, that is because most images in the training set are collected in winter when there is hardly any green plants in scenes. For future work, the raw-format CID dataset can be extended by unprocessing [14] technique and generating raw-format images from other available online datasets. Another limitation is algorithm memory consumption. The batch size is constrained because it takes too much GPU memory to train the proposed model. Therefore training the model can be time-consuming on GPUs with small memory. We are planning to improve network architecture and investigate the possibility of application on mobile devices.

As a future expectation, the model can be further modified by introducing HDRI (High Dynamic Range Imaging). Since our CID dataset is composed of multi-exposed images, we can calculate an HDR image of each scene. Finally, the HDR images can be used as ground truth to participate in the ESN training.

7 References