

# Attention Guided Low-light Image Enhancement with a Large Scale Low-light Simulation Dataset

Feifan Lv · Yu Li · Feng Lu

Received: date / Accepted: date

**Abstract** Low-light image enhancement is challenging in that it needs to consider not only brightness recovery but also complex issues like color distortion and noise, which usually hide in the dark. Simply adjusting the brightness of a low-light image will inevitably amplify those artifacts. To address this difficult problem, this paper proposes a novel end-to-end attention-guided method based on multi-branch convolutional neural network. To this end, we first construct a synthetic dataset with carefully designed low-light simulation strategies. The dataset is much larger and more diverse than existing ones. With the new dataset for training, our method learns two attention maps to guide the brightness enhancement and denoising tasks respectively. The first attention map distinguishes underexposed regions from well lit regions, and the second attention map distinguishes noises from real textures. With their guidance, the proposed multi-branch decomposition-and-fusion enhancement network works in an input adaptive way. Moreover, a reinforcement-net further enhances color and contrast of the output image. Extensive experiments on multiple datasets demonstrate that our method can produce high fidelity enhancement results for low-light images and outperforms the current state-of-the-art methods by a large margin both quantitatively and visually.

---

F. Lv  
Beihang University, Beijing, China.

Y. Li  
Applied Research Center (ARC), Tencent PCG, Shenzhen, China.

F. Lu (Corresponding author)  
Beihang University, Beijing, China  
Peng Cheng Laboratory, Shenzhen, China  
E-mail: lufeng@buaa.edu.cn

**Keywords** Low-light image enhancement · Low-light simulation · Synthetic dataset · Attention guidance · Deep neural network

## 1 Introduction

Images captured in insufficiently illuminated environment usually contain undesired degradations, such as poor visibility, low contrast, unexpected noise, etc. Resolving these degradations and converting low-quality low-light images to normally exposed high-quality images require well developed low-light enhancement techniques. Such a technique has a wide range of applications. For example, it can be used in consumer photography to help the users capture appealing images in the low-light environment. It is also useful for a variety of intelligent systems, *e.g.*, automated driving and video surveillance, to capture high-quality inputs under low-light conditions.

Low-light image enhancement is still a challenging task, since it needs to manipulate color, contrast, brightness and noise simultaneously given the low quality input only. Although numbers of methods have been proposed for this task in recent years, there is still large room for improvement. Figure 1 shows some limitations of existing methods, which follow typical assumptions of histogram equalization (HE) and Retinex theory [44]. HE-based methods aim to increase the contrast by simply stretching the dynamic range of images, while Retinex-based methods recover the contrast by using the estimated illumination map. Mostly, they focus on restoring brightness and contrast and ignore the influences of noise. However, in reality, the noise is inevitable and non-negligible in the low-light images, especially after increasing brightness and contrast.

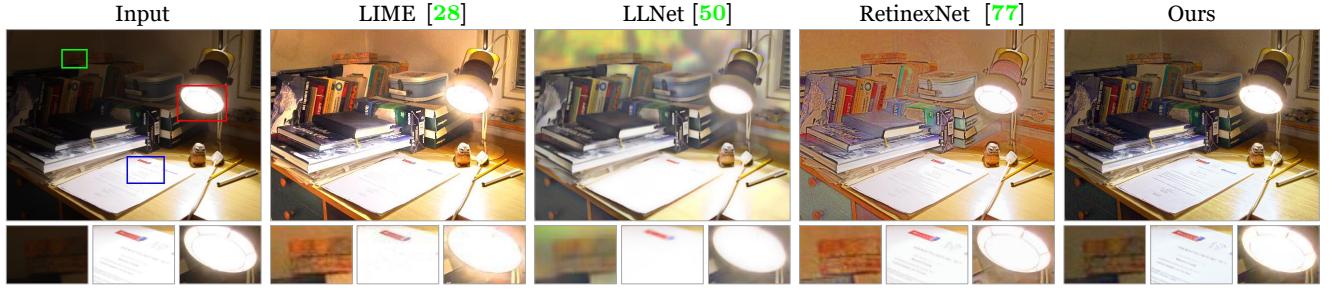


Fig. 1: Low-light enhancement example. Comparing with existing methods, our method can generate results with satisfactory visibility, natural color, and higher contrast.

To suppress the low-light image noise, some methods directly include a denoising process as a separate component in their enhancement pipeline. However, it is dilemma to make a simple cascade of the denoising and enhancement procedures. In particular, applying denoising before enhancement will result in blurring, while applying enhancement before denoising will cause noise amplification. Therefore, in this paper, we propose to model and solve the denoising and low-light enhancement problems simultaneously.

Specifically, this paper proposes an attention-guided enhancement solution that achieves denoising and enhancing simultaneously and effectively. We find that the severity of low brightness/contrast and high image noise show certain spatial distributions related to the underexposed areas. Therefore, the key is to handle the problem in a region-aware adaptive manner. To this end, we propose the under-exposed (ue) attention map to evaluate the degree of underexposure. It guides the method to pay more attention to the underexposed areas in low light enhancement. In addition, based on the ue-attention map, we derive the noise map to guide the denoising according to the joint distribution of exposure and noise intensity. Subsequently, we design a multi-branch CNN to simultaneously achieve low-light enhancement and denoising under the guidance of both maps. In the final step, we add a fully-convolutional network for improving the image contrast, exposure and color as the second enhancement.

The remaining difficulty lies in the lack of large-scale paired low-light image dataset, making it challenging to train an effective network. To address this issue, we propose a low-light image simulation pipeline to synthesize realistic low-light images with well exposed ground truth images. Image contrast and color are also improved to provide good references for our image enhancement step. Following the above ideas, we propose a large-scale low-light image dataset as an efficient benchmark for low-light enhancement researches.

Overall, our contributions are in three folds: 1) We propose a full pipeline for low-light image simulation with high fidelity, based on which we build a new large-scale paired low-light image dataset to support low-light enhancement researches. 2) We propose an attention-guided enhancement method and the corresponding multi-branch network architecture. Guided by the ue-attention map and noise map, the proposed method achieves low-light enhancement and denoising simultaneously and effectively. 3) Comprehensive experiments have been conducted and the experiment results demonstrate that our method outperforms state-of-the-art methods by a large margin.

## 2 Related Work

Image enhancement and denoising have been studied for a long time. In this section, we will briefly overview the most related methods.

**Traditional enhancement methods.** Traditional methods can be mainly divided into two categories. The first category is built upon the histogram equalization (HE) technique. The differences of different HE-based methods are using different additional priors and constraints. In particular, BPDHE [35] tries to preserve image brightness dynamically; Arici *et al.* [4] propose to analyze and penalize the unnatural visual effects for better visual quality; DHECI [54] introduces and uses the differential gray-level histogram; CVC [9] uses the interpixel contextual information; LDR [46] focuses on the layered difference representation of 2D histogram to try to enlarge the gray-level differences between adjacent pixels. These methods expand the dynamic range and focus on improving the contrast of the entire image instead of considering the illumination. They may cause the problem of over- and under-enhancement.

The other category is based on the Retinex theory [44], which assumes that an image is composed of reflection and illumination. Typical methods, *e.g.*,

MSR [40] and SSR [41], try to recover and use the illumination map for low-light image enhancement. Recently, AMSR [47] proposes a weighting strategy based on SSR. NPE [74] balances the enhancement level and image naturalness to avoid over-enhancement. MF [22] processes the illumination map in a multi-scale fashion to improve the local contrast and maintain naturalness. SRIE [23] develops a weighted vibrational model for illumination map estimation. LIME [28] develops a structure-aware smoothing model to estimate the illumination map. BIMEF [82] proposes a dual-exposure fusion algorithm and Ying *et al.* [83] use the camera response model for further enhancement. Mading *et al.* [48] propose a robust Retinex model by considering the noise map for enhancing low-light images accompanied by intensive noise. However, the key to these Retinex-based methods is the estimation of the illumination map, which is hand-crafted and relied on careful parameters tuning. Besides, most of these Retinex-based methods do not consider noise removal and often amplify the noise.

**Learning-based enhancement methods.** Recently, deep learning has achieved great success in the field of low-level image processing [66] and nighttime scenes modeling [57] and understanding [18, 62]. Powerful tools such as end-to-end networks and GANs [25] have been used in image enhancement. LLNet [50] uses the multilayer perception auto-encoder for low-light image enhancement and denoising. HDRNet [24] learns to make local, global, and content-dependent decisions to approximate the desired image transformation. LLCNN [72] and [71] rely on some traditional methods and are not end-to-end solutions to handle brightness/contrast enhancement and denoising simultaneously. MSRNet [68] learns an end-to-end mapping between dark/bright images by using different Gaussian convolution kernels. MBLLEN [51] uses a novel multi-branch low-light enhancement network architecture to learn the mapping from low-light images to normal light ones. Retinex-Net [77] combines the Retinex theory with CNN to estimate the illumination map and enhance the low-light images by adjusting the illumination map. Similarly, KinD [88] designs a similar network by adding a Restoration-Net for noise removal. Ren *et al.* [59] propose a novel hybrid network contains a content stream and a salient edge stream for low-light image enhancement. DeepUPE [73] proposes a network for enhancing underexposed images by estimating an image-to-illumination mapping. However, it does not consider the low-light noise.

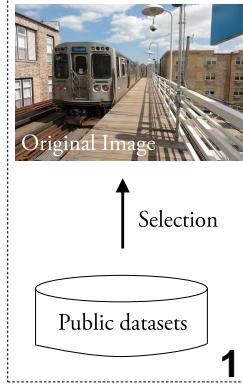
Besides, DPED [36, 70] proposes an end-to-end approach using a composite perceptual error function for translating low-quality mobile phone photos into DSLR-

quality photos. PPCN [34] designs a compact network and combines teacher-student information transfer to reduce computational cost. WESPE [37] proposes a weakly-supervised method to overcome the restrictions on requiring paired images. Also, Chen *et al.* [13] propose an unpaired learning method for image enhancement by improving two-way GANs. As for extremely low-light scenes, SID [10] develops a CNN-based pipeline to directly process raw sensor images. Lv *et al.* [52] propose an enhancement solution by separating the visible and near-infrared signal from a single image and fusing them for high-quality images. Most of these learning-based methods do not explicitly contain the denoising process, and some even rely on traditional denoising methods. However, our approach considers the effects of noise and uses two attention maps to guide the enhancing and denoising process. So, our method is complementary to existing learning-based methods.

**Image denoising methods.** Existing works for image denoising are massive. For Gaussian denoising, BM3D [16] and DnCNN [84] are representatives of the filter-based and deep-learning-based methods. For Poisson denoising, NLPCA [65] combines elements of dictionary learning with sparse patch-based representations of images and employs an adaptation of Principal Component Analysis. Azzari *et al.* [5] propose an iterative algorithm combined with variance-stabilizing transformation (VST) and BM3D filter [16]. DenoiseNet [58] uses a deep convolutional network to calculate the negative noise components, which adds directly to the original noisy image to remove Poisson noise. For Gaussian-Poisson mixed denoising, CBDNet [27] presents a convolutional blind denoising network by incorporating asymmetric learning. It is applicable to real noise images by training on both synthetic and real images. For real-world image denoising, TWSC [78] develops a trilateral weighted sparse coding scheme. Chen *et al.* [11] propose a two-step framework which contains noise distribution estimation using GANs and denoising using CNNs. Directly combining these methods with enhancement methods will result in blurring. To avoid this, our solution performs enhancing and denoising simultaneously.

**Low-light Image Enhancement Datasets.** Several previous datasets are constructed by manually capturing paired low-light and normal-light images. Multiple shootings with different camera configurations or retouching captured images are the two main solutions. The LOL [77] and SID [10] datasets are constructed using the former solution. Images in the LOL dataset are captured in the daytime by controlling the exposure and ISO. Meanwhile, the underexposed images are generated by linear degradation approximately, which may differ from real cases. This will result in performance

## Selecting Original Images (Sec 3.1)



Exposure ↘



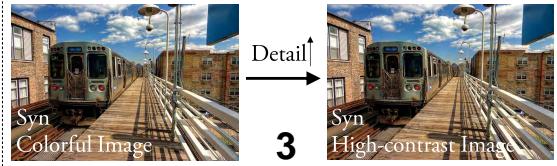
## Generating Realistic Low-light Images (Sec 3.2)

+ Noise ↗



## Generating Reference Images (Sec 3.2)

Contrast ↗



3 Syn High-contrast Image

2 Noise Map

Fig. 2: Pipeline of constructing the proposed low-light simulation dataset. Our method optimally selects normal exposed images from public datasets, performs low-light simulation, and adds noise to synthesize realistic low-light images. Meanwhile, the original normal exposed images are enhanced by exposure correction and contrast/details amplification, so as to generate high-quality reference images. Details can be found in Section 3.

variation in low-light image enhancement (see the result of RetinexNet in Figure 1). The SID dataset is composed of raw sensor data under extremely low-light scenes, which is different from those used in general low-light image enhancement researches. As for the latter solution, the DeepUPE [73] dataset collects 3,000 underexposed images, each with an expert-retouched reference. However, the under-exposed levels of the images are relatively low, which may not cover the heavily low-light scenes. Besides, the SICE [8] dataset collects multi-exposed image sequences and uses the Exposure Fusion methods to construct the reference image under the supervision of human. However, imperfect alignment of image sequences will result in blur and ghosting. Although these datasets have made great contributions to the field of low-light image enhancement, they still show limitations. On one hand, their data amounts are relatively small with respect to the number of images. Since the variation of scenes and light conditions are limited, the trained models may not be generalized well in many cases. On the other hand, due to the lack of annotations, these datasets are difficult to be used for other relevant vision tasks, such as detection and segmentation in the dark.

### 3 Large Scale Low-Light Simulation Dataset

In this paper, we propose an effective low-light simulation method to synthesize low-light images from normal-light images. The purpose is to offer a large diversity in scenes and light conditions which is required by our method and other further researches. Many previous

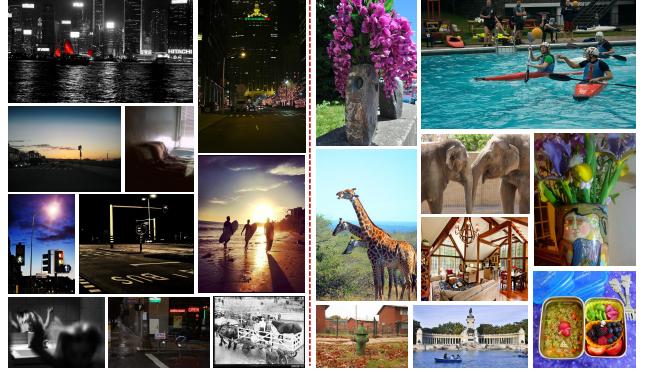


Fig. 3: Samples of large-scale public datasets: (left) low-quality examples, (right) high-quality examples.

works [29, 64] have proven that the synthetic data is an effective alternative to real data in different vision tasks. Using synthetic data allows easy model adaptation for target conditions without requiring additional manual annotations [17, 63]. Similarly, we believe that generating synthetic low-light image datasets from public datasets [6, 21, 26, 49] with rich annotations also has the potential to achieve model adaptation in low-light conditions. The proposed dataset construction pipeline is shown in Figure 2.

#### 3.1 Candidate Image Selection

Our proposed low-light simulation requires high-quality normally exposed images as input, and these images also serve as the reference for low-light enhancement.

Therefore, we need to distinguish such high-quality images from low-quality ones given large-scale public image datasets, as shown in Figure 3. To this end, we propose a candidate image selection method which takes the proper exposure, rich color, blur-free and rich details into account. The selection method contains three steps: darkness estimation, blur estimation and color estimation.

**Darkness estimation.** To select images with sufficient exposure, we first apply over-segmentation [3] and restore the segmentation results. Subsequently, we calculate the mean/variance of the  $V$  component in  $HSV$  color space based on the segmentation results. If the calculated mean/variance is larger than thresholds, we set this segmentation block to be sufficiently exposed. Finally, images with more than 85% sufficiently bright blocks are selected as candidates.

**Blur estimation.** This stage aims to select unblurred images with rich details. Following the same pipeline in [56], we apply the Laplacian edge extraction, calculate the variance among all the output pixels and use a threshold 500 to determine whether this image can be selected.

**Color estimation.** We directly estimate the color according to [30] to select images with rich color. A threshold is set to 500 to eliminate those low-quality, gray-scale or unnatural images.

To ensure diversity, we select 97,030 images from a total of 344,272 images (collected from [6, 21, 26, 49]) based on the above rules to build the dataset. We randomly select 1% of them as the test set which contains 965 images. In this paper, we use the data-balanced subset including 22,656 images as the training set.

### 3.2 Target Image Synthesis

We propose a low-light image simulation method to synthesize realistic low-light images from normal-light images, as shown in Figure 2. This produces an adequate number of paired low/normal light images which are needed for training of learning-based methods.

**Low-light image synthesis.** Low-light images differ from normal images due to two dominant features: low brightness/contrast and the presence of noise. In our low-light image synthesis, we try to fit a transformation to convert the normal image to underexposed low-light image. By analyzing images with different degree of exposure, we find that the combination of linear and gamma transformation can approximate this job well. To verify this, we test on multi-exposure images and use the histogram of  $Y$  channel in  $YCbCr$  color space as the metric. As shown in Figure 4, the synthetic low-light images are approximately the same to

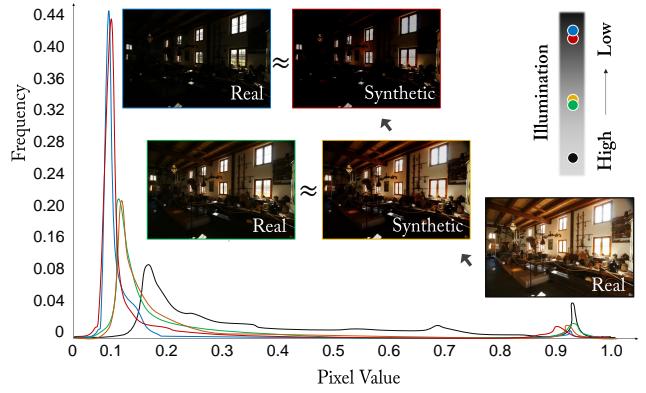


Fig. 4: Verification of the low-light simulation method: visual comparison and the histogram of  $Y$  channel in  $YCbCr$  between synthetic images and real different exposure images.

real low-light images. The low-light image simulation pipeline (without additional noise) can be formulated as:

$$I_{out}^{(i)} = \beta \times (\alpha \times I_{in}^{(i)})^\gamma, i \in \{R, G, B\}. \quad (1)$$

where  $\alpha$  and  $\beta$  are linear transformations, the  $X^\gamma$  means the gamma transformation. The three parameters is sampled from uniform distribution:  $\alpha \sim U(0.9, 1)$ ,  $\beta \sim U(0.5, 1)$ ,  $\gamma \sim U(1.5, 5)$ .

As for the noise, many previous methods fail to consider, while our method takes it into account. In particular, we follow [27, 80] to use the Gaussian-Poisson mixed noise model and take the in-camera image processing pipeline into account to simulate real low-light noise. The noise model can be formulated as:

$$I_{out} = f(M^{-1}(\mathcal{P}(M(f^{-1}(I_{in}))) + N_G)), \quad (2)$$

where  $\mathcal{P}(x)$  represents adding Poisson noise with variance  $\sigma_p^2$ ,  $N_G$  is modeled as AWGN with noise variance  $\sigma_g^2$ ,  $f(x)$  stands for the camera response function,  $M(x)$  is the function that convert RGB images to Bayer images and  $M^{-1}(x)$  is the demosaicing function. We do not consider compression in this paper and the configuration is the same as [27].

**Image contrast amplification.** The high-quality images in our dataset serve as the reference for low-light enhancement. However, directly using them to train an image-to-image regression method may result in low-contrast results (see MBLLEN [51] results in Figure 10). The possible reason caused low-contrast might that some selected images are slightly over-exposed, which will guide enhancement algorithms tend to generate slightly over-exposed results. Besides, the smoothness caused by noise removal also result in low-contrast.

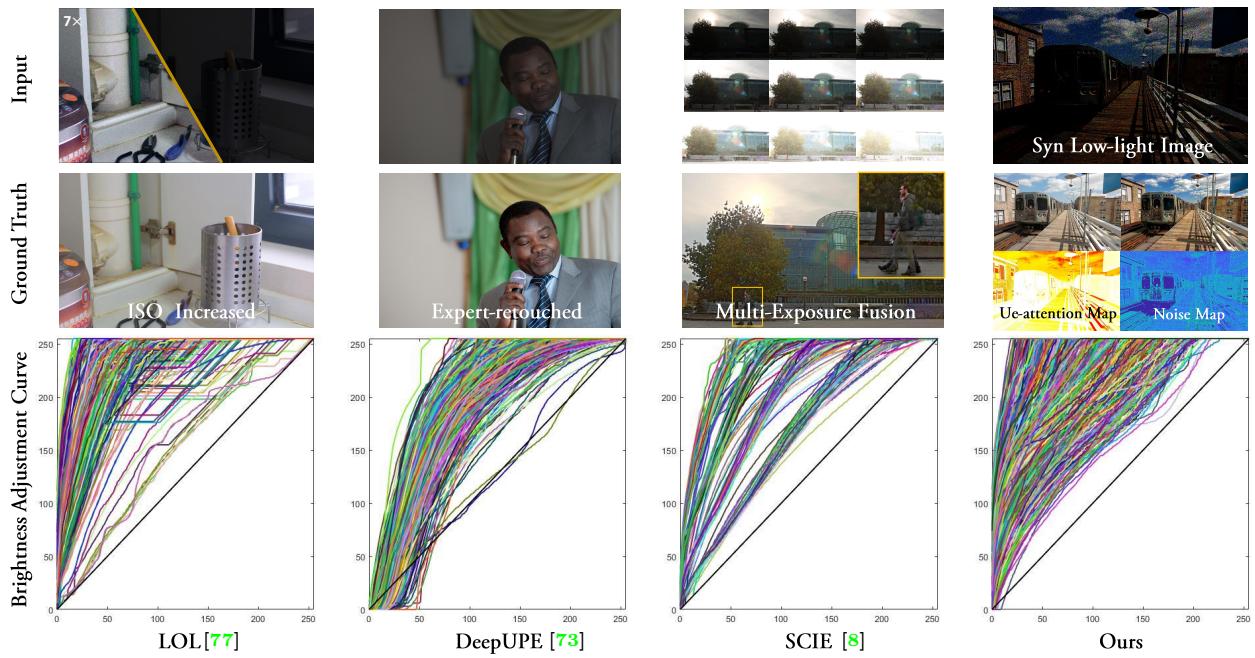


Fig. 5: Comparison with existing paired low-light datasets. **Top:** Example images of different datasets. **Bottom:** The distribution of exposure adjustment curves of different datasets.

To overcome this limitation, we propose a contrast amplification method by synthesizing a new set of high-quality images as the ground truth of our second enhancement step. In particular, we apply exposure fusion to improve the contrast/color and correct the exposure. First, we use gamma transforms to synthesize 10 images with different exposure settings and saturation levels from each original image. Subsequently, we fuse these differently exposed images following the same routine in [53] (the results called colorful images). Finally, we apply image smoothing [79] to further enhance the image details. The final output images called high-contrast images that can be used as ground truth to train a visually better low-light enhancement network.

### 3.3 Comparison of Low-light Enhancement Datasets

There are some existing datasets for low-light image enhancement. However, these datasets still have their own limitations. In this section, we highlight the differences between our synthetic dataset and other low-light image enhancement datasets, to show that our synthetic dataset is a good complement to existing datasets. The characteristics of different datasets are summarized in Table 1.

**Scale and Diversity.** Having a large dataset with diverse scenes and lighting conditions is significant for training a model that can generalize well. Manually col-

Table 1: Comparison with existing low-light enhancement datasets. H(heavy), M(edium) and S(light) means the underexposed level. “MEF” means Multi Expose Fusion methods. “Comp.” means Compatibility, which indicates whether the data acquire method can directly extend to existing public dataset for computer vision problems that have other annotations.

| Dataset      | Level | Source    | Noise | Comp. | Scenes        |
|--------------|-------|-----------|-------|-------|---------------|
| SID [10]     | H     | Camera    | ✓     | ✗     | 424           |
| LOL [77]     | H-M   | Camera    | ✓     | ✗     | 500           |
| SICE [8]     | H-M-S | MEF       | ✗     | ✗     | 589           |
| DeepUPE [73] | M-S   | Retouch   | ✗     | ✗     | 3,000         |
| Ours         | H-M-S | Synthesis | ✓     | ✓     | <b>22,656</b> |

lecting images or editing images as in other datasets is a costly process, which make them hard to acquire data at scale. Therefore, existing datasets are all relatively small in size. In contrast, as our data generation is based on simulation, our method can synthesize paired low-light and normal light images as much as needed for different scenes.

**Low-light Level.** Covering a large range of low-light conditions is another important factor for the generalization capabilities of the trained models. To illustrate the range of different under-exposed levels, we calculate the exposure adjustment curve, which is the transform to the luminance channel of the low-light image (the  $V$  component in HSV color space) to make the

luminance histogram match that of reference ground truth image. The estimated curve can serve as an estimation of the under-exposure levels, that is, steeper change of the curve indicates higher under-exposure levels.

The exposure adjustment curves for all data pair in each dataset are shown in Figure 5. It shows that the LOL [77] dataset contains many heavy low-light images. The DeepUPE [73] dataset mainly covers medium under-exposure levels. As for the SICE [8] dataset, the under-exposure degree is sparse, which is caused by its specific exposure pre-settings. Notice that, in this research, we use the medium exposed images as the ground truth to estimate the curves for SICE [8] dataset. In contrast, our synthetic dataset contains a large variety of under-exposure levels, which is useful for improving the generalization capabilities of our trained models.

**Quality.** For learning-based methods, the quality of images is crucial as it directly decides the performance of training models. SCIE [8] uses multi-exposure image fusion result as the ground truth, which inevitably contains ghosting and blur. SID [10] prolongs exposure time to obtain high-quality night images, which may cause local overexposure and blur. LOL [77] captures paired images by adjusting the ISO, which results in the exposure adjustment being approximately linear. In many cases, simply increasing low-light images linearly can result in good results. As for DeepUPE [73], using retouched images as the ground truth does not have the ability to deal with noise and artifacts. In contrast, our synthetic dataset does not have these problems. Besides, it provides the noise distribution map and exposure map that can be used as supervision to improve the performance of the trained model.

**Compatibility.** Beside making the visual quality more appealing, improving the performance of other vision systems under low-light conditions is another important application for low-light enhancement. However, existing datasets do not contain manual annotations as they are only designed for visual quality enhancement. In contrast, our synthetic dataset can directly use existing public datasets (*e.g.*, COCO [49]) to render low-light images and keep their corresponding annotations, such as bounding boxes for object detection and semantic segmentation masks. Previous works [17, 63] have proven that synthetic data is useful for model adaptation under adverse conditions. Thus, our synthetic dataset also has potential ability to improve the performance of fundamental vision methods to handle low-light conditions, such as object detection and semantic segmentation, etc.

In summary, our synthetic dataset has many advantages over existing datasets. Our synthetic dataset contains high quality paired pixel-aligned images with various scenes, diverse lighting conditions, and different underexposed levels. Moreover, this simulation can be applied to datasets with annotations, which is useful for model adaptation under low-light conditions. Our synthetic dataset is an important complement to existing low-light enhancement datasets.

## 4 Attention-guided Low-light Enhancement

In this section, we introduce the proposed attention-guided enhancement solution, including the network architecture, the loss function and the implementation details.

### 4.1 Network Architecture

We propose a fully convolutional network containing four subnets: an Attention-Net, a Noise-Net, an Enhancement Net and a Reinforce-Net. Figure 6 shows the overall network architecture. The Attention-Net is designed for estimating the illumination to guide the method to pay more attention to the underexposed areas in enhancement. Similarly, the Noise-Net is designed to guide the denoising process. Under their guidance, the multi-branch Enhancement-Net can perform enhancing and denoising simultaneously. The Reinforce-Net is designed for contrast re-enhancement to solve the low-contrast limitation caused by regression. The detailed description is provided below.

**Attention-Net.** We directly adopt U-Net in our implementation. The motivation is to provide a guidance to let Enhancement-Net correctly enhance the underexposed areas and avoid over-enhance the normally exposed areas. The output is an ue-attention map indicating the regional under-exposure level, as shown in Figure 7. The higher the illumination is, the lower ue-attention map values are. The ue-attention map's value range is  $[0, 1]$  and is determined by:

$$A = \frac{|max_c(I) - max_c(\mathcal{F}(I))|}{max_c(I)}, \quad (3)$$

where  $max_c(x)$  returns the maximum value among three color channels,  $I$  is the original bright image and  $\mathcal{F}(I)$  is the synthetic under exposed image.

As shown in Figure 7, the inverted ue-attention map looks somewhat similar to the illumination map of the Retinex model. This infers that our ue-attention map carries important information used by the popular Retinex model. On the other hand, using our inverted ue-attention

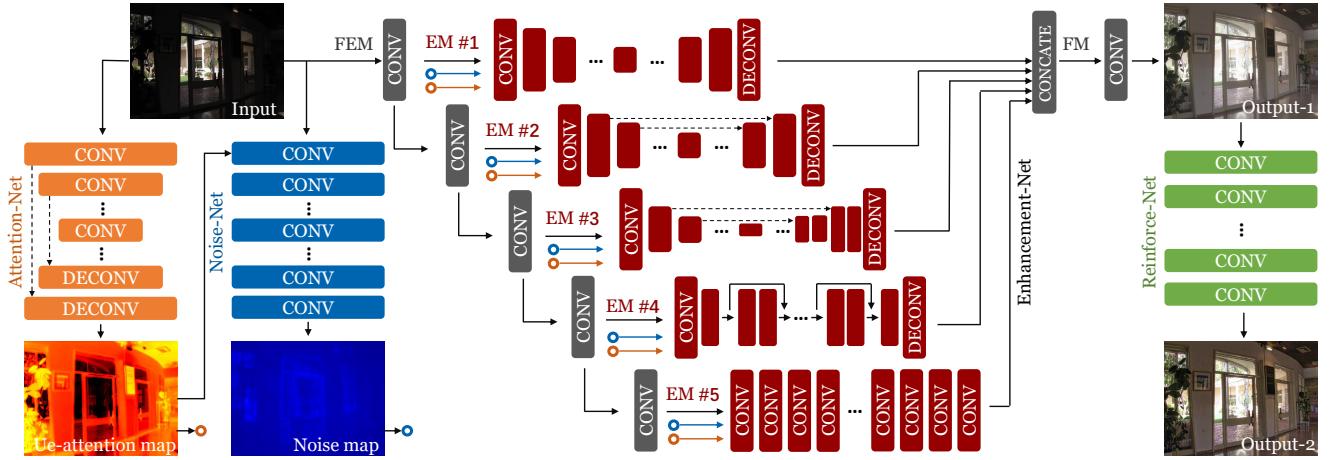


Fig. 6: The proposed network with four subnets. The Attention-Net and Noise-Net are used to estimate the attention of exposure and noise. The Enhancement-Net and Reinforce-Net are corresponding to the two enhancement processes. The core network is the multi-branch Enhancement-Net, which is composed of feature extraction module (FEM), enhancement module (EM) and fusion module (FM). The dashed lines represent skip connections and the circles represent discontinuous connections.

map in Retinex model still cannot ensure satisfactory results. This is because the Retinex-based solution faces difficulties in handling black regions (see black regions in Figure 1) and will result in noise amplification (see LIME results in Figure 11). Therefore, we propose to use the ue-attention map as a guidance for our Enhancement Net introduced later.

**Noise-Net.** The image noise can be easily confused with image textures, causing unwanted blurring effect after applying simple denoising methods. Estimating the noise distribution beforehand and making the denoising adaptive may help reduce such an effect. The noise map's value range is [0, 1] and is determined by:

$$N = \max_c(\frac{|\mathcal{F}_n(I) - \mathcal{F}(I)|}{\mathcal{F}(I)}), \quad (4)$$

where  $\max_c(x)$  returns the maximum value among three color channels,  $\mathcal{F}_n(I)$  is the synthetic low-light image and  $\mathcal{F}(I)$  is the synthetic under exposed image.

Note that the noise distribution is highly related to the distribution of exposure, and thus we propose to use the ue-attention map to help derive a noise map. Under their guidance, the enhancement-net can perform denoising effectively. The Noise-Net is composed of dilated convolutional layers to increase the receptive field, which is conducive to noise estimation.

**Enhancement-Net.** The motivation is to decompose the enhancement problem into several sub-problems of different aspects (such as noise removal, texture preserving, color correction and so on) and solve them respectively to produce the final output via multi-branch



Fig. 7: Comparison between our ue-attention map and the illumination maps used for retinex-based methods. Our ue-attention map can generate similar illumination information with more details.

fusion. It is the core component of the proposed network and it consists of three types of modules: the feature extraction module (FEM), the enhancement module (EM) and the fusion module (FM). **FEM** is a single stream network with several convolutional layers, each of which uses  $3 \times 3$  kernels, stride of 1 and ReLU non-linearity. The output of each layer is both the input to the next layer and also the input to the corresponding subnet of EM. **EMs** are modules following each convolutional layer of the FEM. The input to EM is the output of a certain layer in FEM, and the output size is the same as the input. **FM** accepts the outputs of all EMs to produce the final enhanced image. We concatenate all the outputs from EMs in the color channel dimension and use the  $1 \times 1$  convolution kernel to merge them, which equals to the weighted summation with learnable weights.

We propose five different EM structures. As shown in Figure 6, the design of EM follows U-Net [61] and Res-Net [33] which have been proven effective exten-

sively. In brief, EM-1 is a stack of convolutional and deconvolutional layers with large kernel size. EM-2 and EM-3 has U-Net like structures, and the difference is the skip connection realization and the feature map size. EM-4 has a Res-Net like structure. We remove the Batch-Normalization [39] and use just a few res-blocks to reduce the model parameter. EM-5 is composed of dilated convolutional layers whose output size is the same as the input.

**Reinforce-Net.** The motivation is to overcome the low-contrast drawback and improve the details (see the difference between MBLLEN [51] and ours in Figure 10). Previous research [12] demonstrates the effectiveness of dilated convolution in image processing. Therefore, we use a similar network to improve contrast and details simultaneously.

## 4.2 Loss Function

In order to improve the image quality both qualitatively and quantitatively, we propose a new loss function by further considering the structural information, perceptual information and regional difference of the image. It is expressed as:

$$\mathcal{L} = \omega_a \mathcal{L}_a + \omega_n \mathcal{L}_n + \omega_e \mathcal{L}_e + \omega_r \mathcal{L}_r, \quad (5)$$

where the  $\mathcal{L}_a$ ,  $\mathcal{L}_n$ ,  $\mathcal{L}_e$  and  $\mathcal{L}_r$  represent the loss function of Attention-Net, Noise-Net, Enhancement-Net and Reinforce-Net, and  $\omega_a, \omega_n, \omega_e, \omega_r$  are the corresponding coefficients. The details of the four loss functions are given below.

**Attention-Net loss.** To obtain the correct ue-attention map for guiding the Enhancement-Net, we use the L2 error metric to measure the prediction error as:

$$\mathcal{L}_a = \|\mathcal{F}_a(I) - A\|^2, \quad (6)$$

where  $I$  is the input image,  $\mathcal{F}_a(I)$  and  $A$  are the predicted and expected ue-attention maps.

**Noise-Net loss.** Similarly, we use the L1 error metric to measure the prediction error of the Noise-Net as:

$$\mathcal{L}_n = \|\mathcal{F}_n(I, A') - N\|^1, \quad (7)$$

where  $A' = \mathcal{F}_a(I)$ ,  $\mathcal{F}_n(I, A')$  and  $N$  are the predicted and expected noise maps.

**Enhancement-Net loss.** Due to the low brightness of the image, only using common error metrics such as *mse* or *mae* may cause structure distortion such as

blur effect and artifacts. We design a new loss that consists of four components to improve the visual quality. It is defined as:

$$\mathcal{L}_e = \omega_{eb} \mathcal{L}_{eb} + \omega_{es} \mathcal{L}_{es} + \omega_{ep} \mathcal{L}_{ep} + \omega_{er} \mathcal{L}_{er}, \quad (8)$$

where the  $\mathcal{L}_{eb}$ ,  $\mathcal{L}_{es}$ ,  $\mathcal{L}_{ep}$  and  $\mathcal{L}_{er}$  represent bright loss, structural loss, perceptual loss and regional loss. And  $\omega_{eb}$ ,  $\omega_{es}$ ,  $\omega_{ep}$  and  $\omega_{er}$  are the corresponding coefficients.

The bright loss is designed to ensure that the enhanced results have sufficient brightness. It is defined as:

$$\mathcal{L}_{eb} = \|\mathcal{S}(\mathcal{F}_e(I, A', N') - \tilde{I})\|^1, \quad (9)$$

where  $\mathcal{F}_e(I, A', N')$  and  $\tilde{I}$  are the predicted and expected enhancement images.  $\mathcal{S}$  is defined as:  $\mathcal{S}(x < 0) = -\lambda x$ ,  $\mathcal{S}(x \geq 0) = x$ , s.t.  $\lambda > 1$ .

The structural loss is introduced to preserve the image structure and avoid blurring. We use the well-known image quality assessment algorithm SSIM [76] to build our structure loss. The structural loss is defined as:

$$\mathcal{L}_{es} = 1 - \frac{1}{N} \sum_{p \in img} \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \cdot \frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad (10)$$

where  $\mu_x$  and  $\mu_y$  are pixel value averages,  $\sigma_x^2$  and  $\sigma_y^2$  are variances,  $\sigma_{xy}$  is the covariance, and  $C_1$  and  $C_2$  are constants to prevent the denominator to zero.

The perceptual loss is introduced to use higher-level information to improve the visual quality. We use the well-behaved VGG network [69] as the content extractor [45]. In particular, we define the perceptual loss based on the output of the ReLU activation layers of the pre-trained VGG-19 network. The perceptual loss is defined as follows:

$$\mathcal{L}_{ep} = \frac{1}{w_{ij} h_{ij} c_{ij}} \sum_{x=1}^{w_{ij}} \sum_{y=1}^{h_{ij}} \sum_{z=1}^{c_{ij}} \|\phi_{ij}(I')_{xyz} - \phi_{ij}(\tilde{I})_{xyz}\|, \quad (11)$$

where  $I' = \mathcal{F}_e(I, A', N')$  and  $\tilde{I}$  are the predicted and expected enhancement images, and  $w_{ij}$ ,  $h_{ij}$  and  $c_{ij}$  describe the dimensions of the respective feature maps within the VGG-19 network. Besides,  $\phi_{ij}$  indicates the feature map obtained by  $j$ -th convolution layer in  $i$ -th block of the VGG-19 Network.

For low-light image enhancement, except taking the image as a whole, we should pay more attention to the underexposed regions. We propose the regional loss to

balances the degree of enhancement for different regions. It is defined as:

$$\mathcal{L}_{er} = \|I' \cdot A' - \tilde{I} \cdot A'\|^1 + 1 - ssim(I' \cdot A', \tilde{I} \cdot A') \quad (12)$$

where  $ssim(\cdot)$  represents the image quality assessment algorithm SSIM [76] and  $A'$  is the predicted ue-attention map which is used as the guidance.

**Reinforce-Net loss.** Similar to the Enhancement-Net loss, the Reinforce-Net loss is defined as:

$$\mathcal{L}_r = \omega_{rb}\mathcal{L}_{rb} + \omega_{rs}\mathcal{L}_{rs} + \omega_{rp}\mathcal{L}_{rp}, \quad (13)$$

where  $\mathcal{L}_{rb}$ ,  $\mathcal{L}_{rs}$  and  $\mathcal{L}_{rp}$  represent bright loss, structural loss and perceptual loss, and are the same as  $\mathcal{L}_{rb}$ ,  $\mathcal{L}_{rs}$  and  $\mathcal{L}_{rp}$ . In the experiments, we empirically set  $\lambda = 10$ ,  $\omega_a, \omega_n, \omega_e, \omega_r = \{100, 10, 10, 1\}$ ,  $\omega_{eb}, \omega_{es}, \omega_{ep}, \omega_{er} = \{1, 1, 0.35, 5\}$ ,  $\omega_{rb}, \omega_{rs}, \omega_{rp} = \{1, 1, 0.35\}$ .

### 4.3 Implementation Details

Our implementation is done with Keras [15] and Tensorflow [4]. The proposed network can be quickly converged after being trained for 20 epochs on a Titan-X GPU using the proposed dataset. In order to prevent overfitting, we use random clipping, flipping and rotating for data augmentation. We set the batch-size to 8 and the size of random clipping patches to  $256 \times 256 \times 3$ . The input image values is scaled to  $[0, 1]$ . We use the output of the fourth convolutional layer in the third block of VGG-19 network as the perceptual loss extraction layer.

In the experiment, training is done using the Adam optimizer [42] with parameters of  $\alpha = 0.0002$ ,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$  and  $\epsilon = 10^{-8}$ . We also use the learning rate decay strategy, which reduces the learning rate to 98% before the next epoch. At the same time, we reduce the learning rate to 50% when the loss metric has stopped improving.

## 5 Experimental Evaluation

We compare our method with existing methods through extensive experiments. We use the publicly-available codes with recommended parameter settings. In quantitative comparison, we used PSNR and SSIM [76], along with some recently proposed metrics *Average Brightness* (AB) [14], *Visual Information Fidelity* (VIF) [67], *Lightness Order Error* (LOE) [82], *Tone Mapped Image Quality Index* (TMQI) [81] and *Learned Perceptual Image Patch Similarity Metric* (LPIPS)[86]. For all metrics higher number means better, except LPIPS, LOE

Table 2: Quantitative comparison of synthetic low-light image (without additional noise) enhancement. “ $\uparrow$ ” indicates the higher the better, “ $\downarrow$ ” indicates the lower the better, “ $\Downarrow$ ” indicates the lower absolute value the better.

|              | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | VIF $\uparrow$ | LOE $\downarrow$ | TMQI $\uparrow$ | AB $\Downarrow$ |
|--------------|-----------------|-----------------|--------------------|----------------|------------------|-----------------|-----------------|
| Input        | 11.99           | 0.45            | 0.26               | 0.33           | 677.85           | 0.80            | -59.22          |
| BIMEF [82]   | 18.28           | 0.76            | <b>0.11</b>        | 0.49           | 550.20           | 0.85            | -28.06          |
| LIME [28]    | 15.80           | 0.68            | 0.20               | 0.48           | 1121.17          | 0.80            | <b>-2.46</b>    |
| MSRCR [40]   | 14.87           | 0.72            | 0.15               | 0.52           | 1249.24          | 0.82            | 35.07           |
| MF [22]      | 15.89           | 0.68            | 0.18               | 0.44           | 766.00           | 0.83            | -36.88          |
| SRIE [23]    | 13.83           | 0.56            | 0.21               | 0.37           | 787.42           | 0.82            | -47.86          |
| Dong [20]    | 15.37           | 0.65            | 0.22               | 0.35           | 1228.49          | 0.81            | -33.80          |
| NPE [74]     | 14.93           | 0.66            | 0.18               | 0.42           | 875.15           | 0.83            | -41.35          |
| DHECI [54]   | 18.13           | 0.76            | 0.17               | 0.39           | 547.12           | 0.87            | -17.37          |
| BDPHE [35]   | 13.62           | 0.60            | 0.24               | 0.34           | 609.89           | 0.82            | -47.82          |
| HE           | 17.88           | 0.76            | 0.18               | 0.47           | 596.67           | <b>0.88</b>     | 19.24           |
| Ying [83]    | 19.21           | <b>0.80</b>     | <b>0.11</b>        | 0.56           | 778.67           | 0.83            | -9.28           |
| WAHE [4]     | 15.46           | 0.65            | 0.18               | 0.44           | 564.83           | 0.84            | -39.38          |
| JED [60]     | 16.11           | 0.65            | 0.21               | 0.41           | 1212.66          | 0.82            | -25.95          |
| Robust [48]  | 16.83           | 0.69            | 0.20               | 0.47           | 1052.22          | 0.82            | -22.09          |
| LLNet [50]   | 20.11           | <b>0.80</b>     | 0.39               | 0.40           | 1088.43          | 0.87            | 4.30            |
| DeepUPE [73] | 16.55           | 0.64            | 0.17               | 0.55           | 516.47           | 0.84            | -30.48          |
| GLADNet [75] | <b>24.57</b>    | <b>0.90</b>     | <b>0.09</b>        | <b>0.62</b>    | <b>513.18</b>    | <b>0.91</b>     | 5.52            |
| MBLLEN [51]  | <b>24.21</b>    | <b>0.90</b>     | <b>0.08</b>        | <b>0.63</b>    | <b>536.75</b>    | <b>0.91</b>     | <b>-3.66</b>    |
| Ours         | <b>25.24</b>    | <b>0.94</b>     | <b>0.08</b>        | <b>0.67</b>    | <b>495.48</b>    | <b>0.93</b>     | <b>2.04</b>     |

Table 3: Quantitative comparison of synthetic low-light images (with additional noise) enhancement. “ $\uparrow$ ” indicates the higher the better, “ $\downarrow$ ” indicates the lower the better, “ $\Downarrow$ ” indicates the lower absolute value the better.

|              | PSNR $\uparrow$ | SSIM $\uparrow$ | LPIPS $\downarrow$ | VIF $\uparrow$ | LOE $\downarrow$ | TMQI $\uparrow$ | AB $\Downarrow$ |
|--------------|-----------------|-----------------|--------------------|----------------|------------------|-----------------|-----------------|
| Input        | 11.23           | 0.37            | 0.41               | 0.23           | 925.06           | 0.77            | -65.32          |
| BIMEF [82]   | 16.57           | 0.64            | 0.32               | <b>0.28</b>    | 978.96           | 0.83            | -32.65          |
| LIME [28]    | 14.79           | 0.59            | 0.34               | 0.26           | 1462.64          | 0.79            | -7.39           |
| MSRCR [40]   | 14.83           | 0.62            | 0.34               | 0.27           | 1559.05          | 0.84            | 30.98           |
| MF [22]      | 15.29           | 0.59            | 0.33               | 0.26           | 1095.33          | 0.82            | -37.46          |
| SRIE [23]    | 13.10           | 0.48            | 0.37               | 0.25           | 1095.30          | 0.80            | -52.53          |
| Dong [20]    | 14.69           | 0.56            | 0.35               | 0.21           | 1592.27          | 0.79            | -33.99          |
| NPE [74]     | 14.56           | 0.58            | 0.33               | 0.25           | 1302.10          | 0.82            | -41.17          |
| DHECI [54]   | 16.57           | 0.61            | 0.37               | 0.23           | 924.78           | 0.86            | -15.20          |
| BDPHE [35]   | 12.60           | 0.48            | 0.38               | 0.23           | 925.56           | 0.79            | -54.66          |
| HE           | 16.65           | 0.64            | 0.36               | 0.26           | 1036.22          | 0.87            | 20.21           |
| Ying [83]    | 17.18           | 0.67            | 0.31               | <b>0.28</b>    | 1152.94          | 0.83            | -13.97          |
| WAHE [4]     | 13.97           | 0.52            | 0.36               | 0.27           | 935.21           | 0.81            | -46.87          |
| JED [60]     | 13.70           | 0.48            | 0.46               | 0.22           | 1531.84          | 0.77            | -33.11          |
| Robust [48]  | 14.03           | 0.50            | 0.46               | 0.23           | 1448.03          | 0.77            | -29.09          |
| LLNet [50]   | 18.40           | 0.69            | 0.56               | 0.26           | 1168.75          | 0.85            | -5.25           |
| DeepUPE [73] | 14.94           | 0.53            | 0.35               | 0.25           | 1084.08          | 0.81            | -36.53          |
| GLADNet [75] | <b>19.86</b>    | <b>0.76</b>     | <b>0.19</b>        | <b>0.30</b>    | <b>796.87</b>    | <b>0.88</b>     | <b>5.09</b>     |
| MBLLEN [51]  | <b>19.27</b>    | <b>0.73</b>     | <b>0.23</b>        | <b>0.30</b>    | <b>864.57</b>    | <b>0.89</b>     | <b>-4.87</b>    |
| Ours         | <b>20.84</b>    | <b>0.82</b>     | <b>0.17</b>        | <b>0.33</b>    | <b>785.64</b>    | <b>0.91</b>     | <b>4.36</b>     |

and AB. Note that in the tables below, red, green and blue colors indicate the best, second, and third place results, respectively.

Our experiment is organized as following. First, we make qualitative and quantitative comparisons based on our synthetic dataset and two public-available real low-light datasets. Second, we make visual comparisons with state-of-the-art methods on natural low-light images and provide a user study. We also show the robustness of our method and the benefit to some high-level tasks. Finally, we provide an ablation study to evaluate the effect of different elements and discuss unsatisfying cases.



Fig. 8: Visual comparison on synthetic low-light images. We fine tune the GLADNet [75] using our synthetic datasets. Please zoom in for a better view.

### 5.1 Experiments on Synthetic Datasets

**Direct comparison.** We compare our method with state-of-the-art methods on our synthetic dataset. Since most methods do not have the ability to remove noise, we combine them with the state-of-the-art denoising method CBDNet [27] to produce the final comparison results. We fine tune the GLADNet [75] and LLNet [50] for fair comparison. Quantitative comparison results are shown in Table 2 and Table 3. Our result significantly outperforms other methods in all quality metrics, which fully demonstrates the superiority of our approach.

Representative results are visually shown in Figure 8. By checking the details, it is clear that our method achieves better visual effects, including good brightness/contrast and less artifacts. Please zoom in to compare the details.

**Efficiency comparison.** In addition to the result quality, efficiency is also an important metric to algorithms. In order to demonstrate the superiority of our method, we use 10 HD images with size  $1920 \times 1080$  as the benchmark to test running time. In order to more intuitively demonstrate the relationship between performance and efficiency, we show Figure 9. Our method performs well in terms of both quality and efficiency. Notice that, JED [60] and Robust [48] need large computational resources, which will cause out-of-memory problem when processing large images. Due to the MLP

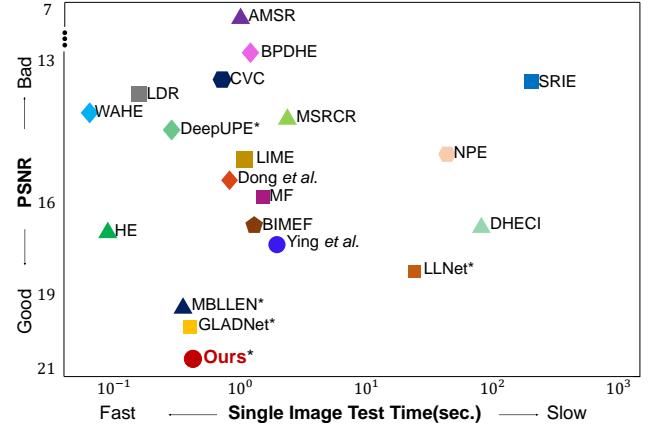


Fig. 9: Runtime and performance comparison of different enhancement methods. Test machine is a PC with Intel i5-8400 CPU, 16 GB memory and NVIDIA Titan-Xp GPU. “\*” represents using GPU.

architecture, LLNet [50] needs to enhance large images one patch by one patch, which will limit its efficiency.

### 5.2 Experiments on Real Datasets

Besides synthetic datasets, our method also performs well on real low-light image datasets. We evaluate the performance based on two public-available real low-

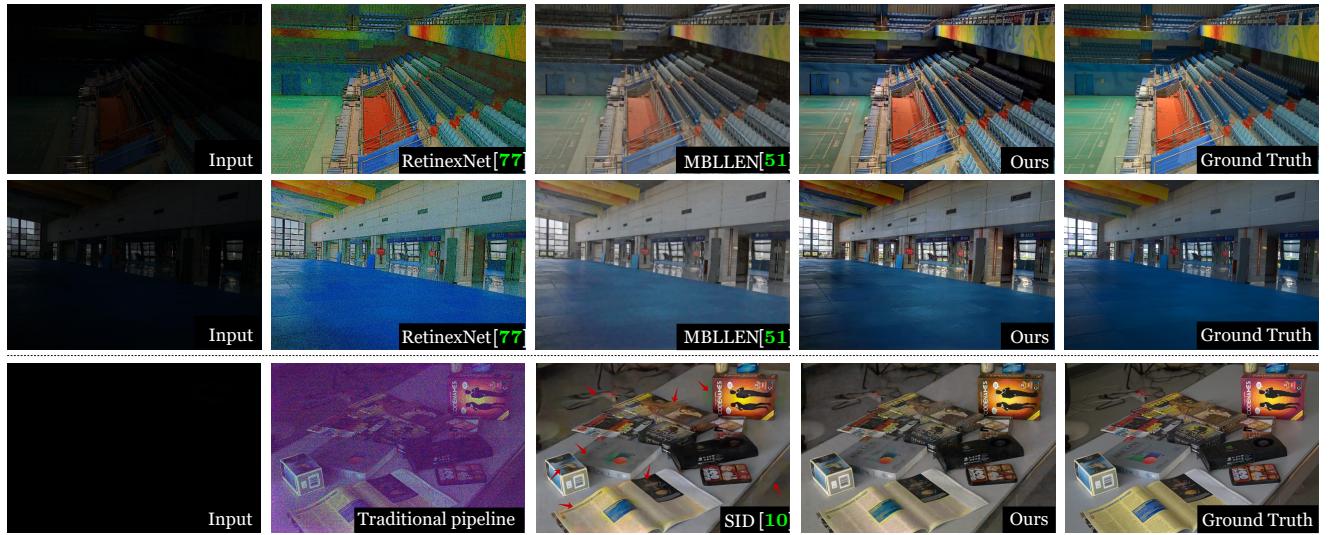


Fig. 10: Visual comparison on the LOL dataset (row 1 and 2) and the SID dataset (row 3). Please zoom in for a better view.

light datasets and show the visual comparison on challenging images.

**LOL dataset.** This dataset is captured by control the exposure and ISO in the daytime. We fine-tune our model using this dataset to compare with RetinexNet [10], which is trained on the LOL dataset. In addition, we replace the Enhancement-Net by a standard U-Net to build a lightweight version. Following PPCN [34], we also adopt knowledge transfer to further promote its performance. Quantitative comparison is shown in Table 4. For both quality and efficiency comparisons, our method performs better, manifesting that our method effectively learns the adjustment and restoration. Visual comparison is shown in Figure 10. Compared with RetinexNet [77] and MBLLEN [51], our results with clear details, better contrast, normal brightness and natural white balance.

**SID dataset.** This dataset contains raw short-exposure images with corresponding long-exposure reference images and is benchmarking single-image processing of extremely low-light raw images. Due to the larger bit depth, raw images are more suitable for extremely low-light scenes compared with rgb images. Different from traditional pipelines, SID [10] develop a pipeline based on an end-to-end network and achieve excellent results. Need to notice that, processing low-light raw images is a related but not identical problem. However, to prove the ability of our multi-branch network, we use the same configuration except that the network is replaced by our Enhancement-Net. Quantitative comparison is shown in Table 4. Our model is lightweight and more efficient, but achieves comparable enhancement quality. In addi-

Table 4: Quantitative comparison between our method and state-of-the-arts on the LOL dataset and the SID dataset. “ours-1” means the result of the Enhancement-Net, “ours-2” means the result of the Reinforce-Net.

| Method                 | PSNR         | SSIM        | LPIPS       | Time        | Params       |
|------------------------|--------------|-------------|-------------|-------------|--------------|
| RetinexNet [77]        | 16.77        | 0.56        | 0.47        | 0.06        | 0.44M        |
| RetinexNet [77] + BM3D | 17.91        | 0.73        | 0.22        | 2.75        | 0.44M        |
| MBLLEN [51]            | 18.56        | 0.75        | 0.19        | <b>0.05</b> | 0.31M        |
| Ours-lightweight-1     | 19.08        | 0.74        | 0.17        | <b>0.05</b> | <b>0.21M</b> |
| Ours-lightweight-2     | 18.79        | 0.77        | 0.21        | <b>0.05</b> | 0.25M        |
| Ours-1                 | <b>20.24</b> | 0.79        | <b>0.14</b> | 0.06        | 0.88M        |
| Ours-2                 | 19.48        | <b>0.81</b> | 0.16        | 0.06        | 0.92M        |
| SID [10]               | <b>28.88</b> | <b>0.79</b> | <b>0.36</b> | 0.51        | 7.76M        |
| Ours                   | 27.96        | 0.77        | <b>0.36</b> | <b>0.48</b> | <b>0.88M</b> |

tion, our results have better visual effects as shown in Figure 10.

### 5.3 Experiments on Real Images

In this section, we evaluate our method on real low-light images, including natural, monochrome and game scenes. We also show the benefit to object detection and semantic segmentation under low-light environment by directly using our method as the pre-processing.

**Natural low-light images.** We first compare our method with state-of-the-art methods on natural low-light images and the representative visual comparison results are shown in Figure 11. Our method surpasses other methods in two key aspects. On the one hand, our method can restore vivid and natural color to make the enhancement results more realistic. In contrast, Retinex-based methods (such as RetinexNet [77] and LIME [28])

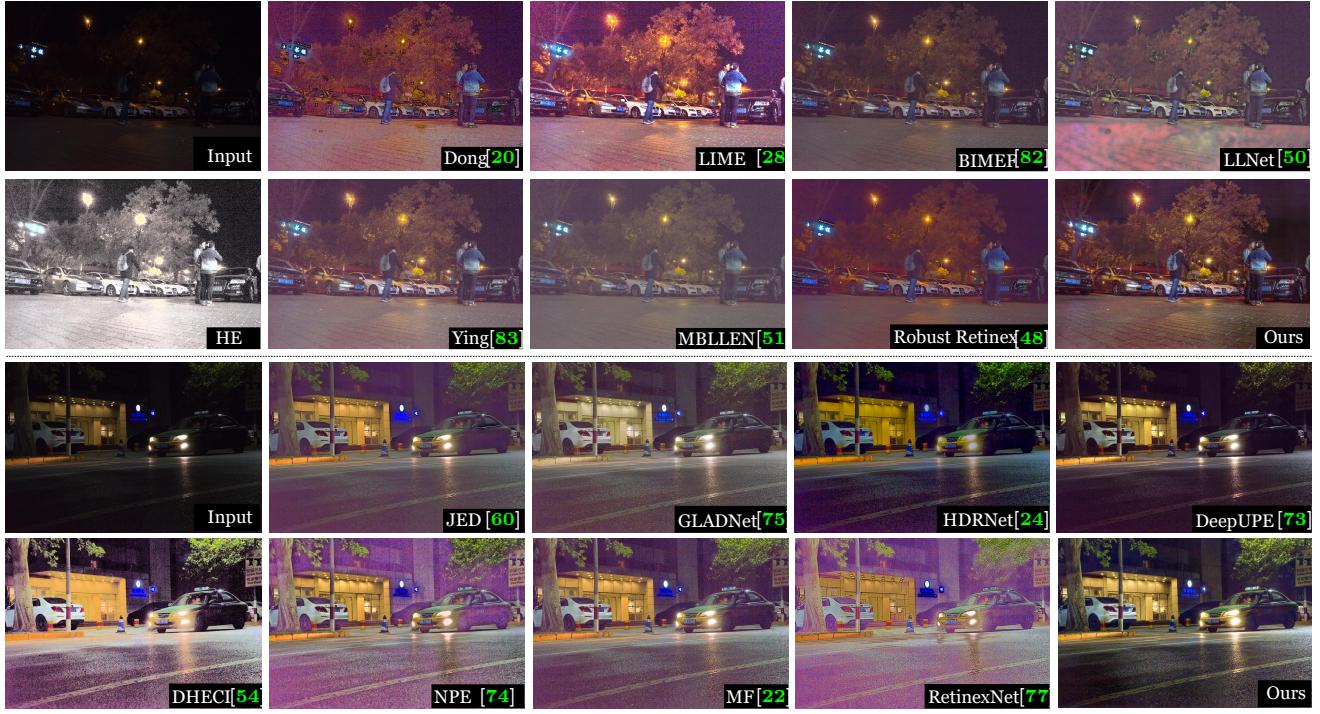


Fig. 11: Visual comparison of real low-light images, which are taken at night. Please zoom in for a better view.

will cause different degrees of color distortion. On the other hand, our method is able to recover better contrast and more details. This improvement is especially evident when compared with LLNet [50], BIMEF [82] and MBLLEN [51].

**User study.** We invite 100 participants to attend a user study to test the subjective preference of low-light image enhancement methods. We randomly select 20 natural low-light image cases and enhance them using five representative methods. For each case, the input data and the five enhanced results will be shown to the participants at the same time. We then ask the participants to rank the quality of the five enhancements from 1 (best) to 5 (worst) in terms of recovery of brightness, contrast, and color. We also provide zoom-in function to let participants to check details like texture and noises controls. The other four methods used besides ours in this study are DHECI [54], DeepUPE [73], LIME [28] and Robust [48].

Figure 12 shows the rating distribution of the user study. Our method receives more “best” ratings, which shows that our results are more preferred by human subjects.

**Generalization study.** To prove the robustness of our method, we directly apply our trained model to enhance some specific types of low-light scenes (such as monochrome surveillance and game night scenes) that

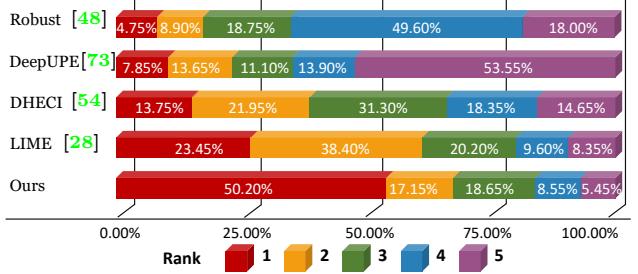


Fig. 12: Rating distribution of the user study.

are unseen in the training dataset. Figure 13 shows the enhancement results. The results demonstrate that our method is robust and effective for general low-light image enhancement tasks. Besides, we also show that our approach is beneficial to some high-level tasks in low-light scenes, such as object detection and instance segmentation, as shown in Figure 14. The performance of Mask-RCNN [2, 31] has been improved a lot by using our method in a pre-processing stage without any fine-tuning. Besides, we have tested end-to-end training our multi-branch network for many other low-level computer vision tasks and demonstrated the effectiveness. Visual examples on denoising, dehazing, deblurring, etc., are shown in Figure 16.



Fig. 13: Generalizing our method to enhance (upper) monochrome surveillance scenes and (bottom) nighttime game scenes.

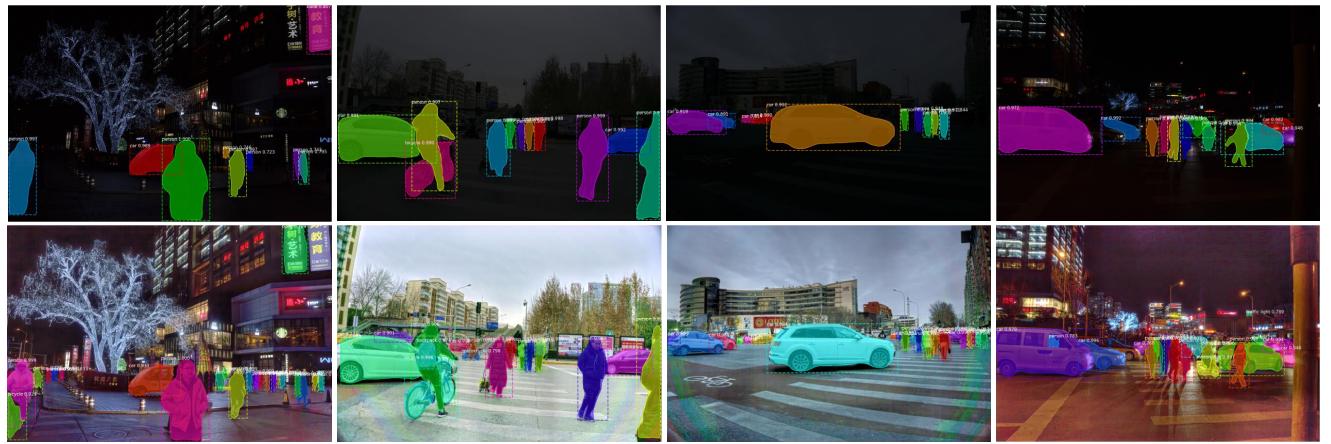


Fig. 14: After processing the low-light scene (upper row) with our method, the performance of both object detection and instance segmentation are greatly improved (bottom row).

#### 5.4 Ablation Study

In this section, we quantitatively evaluate the effectiveness of different components in our method based on our synthetic low-light dataset. Table 5 reports the accuracy of the presented change in terms of PSNR and SSIM [76]. Note that the Reinforce-Net is not considered in this study.

**Loss functions.** We mainly evaluate the loss function of the Enhancement-Net, as shown in Table 5 (row 2-5). We use  $mse$  as the naive loss function under condition 2. The results show that the quality of enhancement is improving by containing more loss components.

**Network structures.** As shown in Table 5 (row 6-7), we evaluate the effectiveness of different network components. Similar to the loss function, the results demonstrate that more components of our network will result in better performance.

**Number of branches.** We analyze the effect of different branch numbers (model size) on the network

Table 5: Ablation study. This table reports the performance under each condition based on the synthetic low-light dataset. In this table, "w/o" means without.

| Condition   | PSNR         | SSIM        |
|---|--------------|-------------|
| 1. default configuration  | <b>20.84</b> | <b>0.82</b> |
| 2. w/o $\mathcal{L}_{eb}$ , w/o $\mathcal{L}_{es}$ , w/o $\mathcal{L}_{ep}$ , w/o $\mathcal{L}_{er}$    | 19.36        | 0.73        |
| 3. with $\mathcal{L}_{eb}$ , w/o $\mathcal{L}_{es}$ , w/o $\mathcal{L}_{ep}$ , w/o $\mathcal{L}_{er}$   | 20.01        | 0.76        |
| 4. with $\mathcal{L}_{eb}$ , with $\mathcal{L}_{es}$ , w/o $\mathcal{L}_{ep}$ , w/o $\mathcal{L}_{er}$  | 19.92        | 0.78        |
| 5. with $\mathcal{L}_{eb}$ , with $\mathcal{L}_{es}$ , with $\mathcal{L}_{ep}$ , w/o $\mathcal{L}_{er}$ | 20.58        | 0.81        |
| 6. w/o Attention-Net, w/o Noise-Net   | 19.12        | 0.71        |
| 7. with Attention-Net, w/o Noise-Net  | 20.66        | 0.80        |
| 8. branch number $\times 1$ (5)   | 20.66        | 0.79        |
| 9. branch number $\times 3$ (15)  | 20.83        | <b>0.82</b> |

performance, as shown in Table 5 (row 8-9). Obviously, the increase of model size will not always improve performance, so we set 10 branches as the default configuration.



Fig. 15: This image has totally dark region where textures are lost and heavy compression. These cause issues in the enhancement result.

### 5.5 Unsatisfying Cases

Figure 15 presents a case where our method performs not perfectly. Our method fails to recover the face details on the top image, as some parts of the face are totally dark. Another issue is the blocking artifacts due to heavy image compression.

## 6 Conclusion

This paper proposes an attention-guided enhancement solution for low-light image enhancement. We design a multi-branch network to handle enhance the brightness and handle the noise simultaneously. The key is to use the proposed ue-attention map and noise map to guide the enhancement in a region adaptive manner. We also propose a low-light image simulation pipeline and build a large-scale low-light enhancement benchmark dataset for model training and evaluation. Extensive experiments demonstrate that our solution outperforms state-of-the-art methods by a large margin. As for future direction, extending the proposed method to low-light video enhancement is of our interest.

## References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al.: Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 (2016)
2. Abdulla, W.: Mask r-cnn for object detection and instance segmentation on keras and tensorflow. [https://github.com/matterport/Mask\\_RCNN](https://github.com/matterport/Mask_RCNN) (2017)
3. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **34**(11), 2274–2282 (2012)
4. Arici, T., Dikbas, S., Altunbasak, Y.: A histogram modification framework and its application for image contrast enhancement. IEEE Transactions on image processing (TIP) **18**(9), 1921–1935 (2009)
5. Azzari, L., Foi, A.: Variance stabilization for noisy+ estimate combination in iterative poisson denoising. IEEE signal processing letters **23**(8), 1086–1090 (2016)
6. Bileschi, S.M.: Streetscenes: Towards scene understanding in still images. Tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE (2006)
7. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: An end-to-end system for single image haze removal. IEEE Transactions on Image Processing (TIP) **25**(11), 5187–5198 (2016)
8. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing (TIP) **27**(4), 2049–2062 (2018)
9. Celik, T., Tjahjadi, T.: Contextual and variational contrast enhancement. IEEE Transactions on Image Processing (TIP) **20**(12), 3431–3441 (2011)
10. Chen, C., Chen, Q.C., Xu, J., Koltun, V.: Learning to see in the dark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
11. Chen, J., Chen, J., Chao, H., Yang, M.: Image blind denoising with generative adversarial network based noise modeling. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
12. Chen, Q., xu, J., Koltun, V.: Fast image processing with fully-convolutional networks. IEEE International Conference on Computer Vision (ICCV) (2017)
13. Chen, Y.S., Wang, Y.C., Kao, M.H., Chuang, Y.Y.: Deep photo enhancer: Unpaired learning for image enhance-

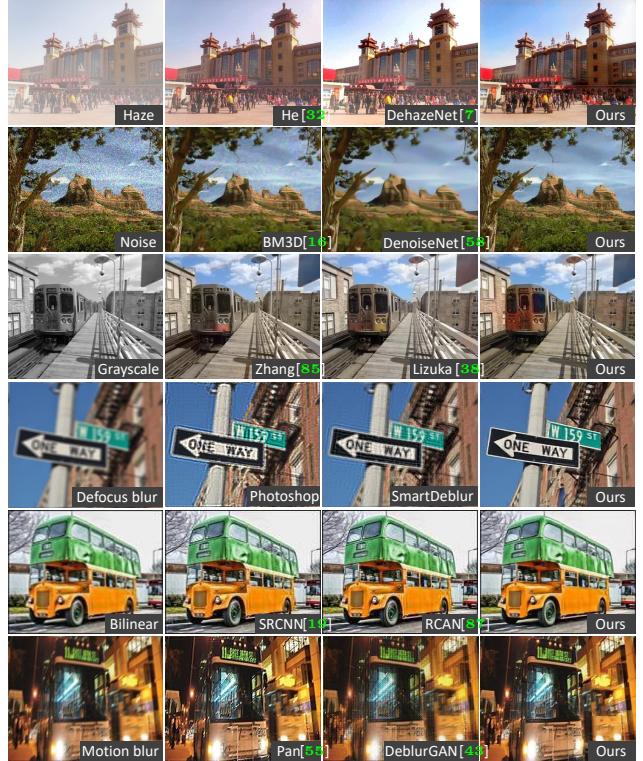


Fig. 16: Visual comparison of several low-level vision tasks. From top to bottom: dehazing, denoising, colorization, defocus deblurring, super-resolution ( $\times 2$ ) and motion deblurring.

- ment from photographs with gans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
14. Chen, Z., Abidi, B.R., Page, D.L., Abidi, M.A.: Gray-level grouping (glg): an automatic method for optimized image contrast enhancement-part i: the basic method. *IEEE transactions on image processing (TIP)* **15**(8), 2290–2302 (2006)
  15. Chollet, F., et al.: Keras. <https://github.com/keras-team/keras> (2015)
  16. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising with block-matching and 3d filtering. In: Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, vol. 6064, p. 606414 (2006)
  17. Dai, D., Sakaridis, C., Hecker, S., Van Gool, L.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. *International Journal of Computer Vision (IJCV)* (2019)
  18. Dai, D., Van Gool, L.: Dark model adaptation: Semantic image segmentation from daytime to nighttime. In: IEEE International Conference on Intelligent Transportation Systems (2018)
  19. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: European conference on computer vision (ECCV) (2014)
  20. Dong, X., Wang, G., Pang, Y., Li, W., Wen, J., Meng, W., Lu, Y.: Fast efficient algorithm for enhancement of low lighting video. In: IEEE International Conference on Multimedia and Expo (ICME) (2011)
  21. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. *International journal of computer vision (IJCV)* **88**(2), 303–338 (2010)
  22. Fu, X., Zeng, D., Huang, Y., Liao, Y., Ding, X., Paisley, J.: A fusion-based enhancing method for weakly illuminated images. *Signal Processing* **129**, 82–96 (2016)
  23. Fu, X., Zeng, D., Huang, Y., Zhang, X.P., Ding, X.: A weighted variational model for simultaneous reflectance and illumination estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
  24. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)* **36**(4), 118 (2017)
  25. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems (NIPS) (2014)
  26. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In: Int. Workshop OntoImage, vol. 5 (2006)
  27. Guo, S., Yan, Z., Zhang, K., Zuo, W., Zhang, L.: Toward convolutional blind denoising of real photographs. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019)
  28. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing (TIP)* **26**(2), 982–993 (2017)
  29. Hahner, M., Dai, D., Sakaridis, C., Zaech, J.N., Van Gool, L.: Semantic understanding of foggy scenes with purely synthetic data. In: IEEE International Conference on Intelligent Transportation Systems (2019)
  30. Hasler, D., Suesstrunk, S.E.: Measuring colorfulness in natural images. In: Human vision and electronic imaging VIII, vol. 5007, pp. 87–96 (2003)
  31. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision (ICCV), pp. 2961–2969 (2017)
  32. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **33**(12), 2341–2353 (2011)
  33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition (CVPR) (2016)
  34. Hui, Z., Wang, X., Deng, L., Gao, X.: Perception-preserving convolutional networks for image enhancement on smartphones. In: European Conference on Computer Vision Workshop (ECCVW) (2018)
  35. Ibrahim, H., Kong, N.S.P.: Brightness preserving dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics* **53**(4), 1752–1758 (2007)
  36. Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: Dslr-quality photos on mobile devices with deep convolutional networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
  37. Ignatov, A., Kobyshev, N., Timofte, R., Vanhoey, K., Van Gool, L.: Wespe: weakly supervised photo enhancer for digital cameras. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2018)
  38. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)* **35**(4), 110 (2016)
  39. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (ICML) (2015)
  40. Jobson, D.J., Rahman, Z.u., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing (TIP)* **6**(7), 965–976 (1997)
  41. Jobson, D.J., Rahman, Z.u., Woodell, G.A.: Properties and performance of a center/surround retinex. *IEEE Transactions on Image processing (TIP)* **6**(3), 451–462 (1997)
  42. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
  43. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., Matas, J.: Deblurgan: Blind motion deblurring using conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
  44. Land, E.H.: The retinex theory of color vision. *Scientific American* **237**(6), 108–129 (1977)
  45. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. *IEEE conference on computer vision and pattern recognition (CVPR)* (2017)
  46. Lee, C., Lee, C., Kim, C.S.: Contrast enhancement based on layered difference representation of 2d histograms. *IEEE transactions on image processing (TIP)* **22**(12), 5372–5384 (2013)
  47. Lee, C.H., Shih, J.L., Lien, C.C., Han, C.C.: Adaptive multiscale retinex for image contrast enhancement. In: Signal-Image Technology & Internet-Based Systems (SITIS) (2013)

48. Li, M., Liu, J., Yang, W., Sun, X., Guo, Z.: Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing (TIP)* **27**(6), 2828–2841 (2018)
49. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision (ECCV) (2014)
50. Lore, K.G., Akintayo, A., Sarkar, S.: Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition (PR)* **61**, 650–662 (2017)
51. Lv, F., Lu, F., Wu, J., Lim, C.: Mbllen: Low-light image/video enhancement using cnns. *British Machine Vision Conference (BMVC)* (2018)
52. Lv, F., Zheng, Y., Li, Y., Lu, F.: An integrated enhancement solution for 24-hour colorful imaging. In: AAAI Conference on Artificial Intelligence (AAAI) (2020)
53. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: Computer Graphics and Applications, pp. 382–390 (2007)
54. Nakai, K., Hoshi, Y., Taguchi, A.: Color image contrast enhacement method based on differential intensity/saturation gray-levels histograms. In: Intelligent Signal Processing and Communications Systems (ISPACS) (2013)
55. Pan, J., Sun, D., Pfister, H., Yang, M.H.: Blind image deblurring using dark channel prior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
56. Pech-Pacheco, J.L., Cristóbal, G., Chamorro-Martínez, J., Fernández-Valdivia, J.: Diatom autofocusing in brightfield microscopy: a comparative study. In: *Pattern Recognition (PR)*, vol. 3, pp. 314–317 (2000)
57. Radenovic, F., Schonberger, J.L., Ji, D., Frahm, J.M., Chum, O., Matas, J.: From dusk till dawn: Modeling in the dark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5488–5496 (2016)
58. Remez, T., Litany, O., Giryes, R., Bronstein, A.M.: Deep convolutional denoising of low-light images. arXiv preprint arXiv:1701.01687 (2017)
59. Ren, W., Liu, S., Ma, L., Xu, Q., Xu, X., Cao, X., Du, J., Yang, M.H.: Low-light image enhancement via a deep hybrid network. *IEEE Transactions on Image Processing (TIP)* (2019)
60. Ren, X., Li, M., Cheng, W.H., Liu, J.: Joint enhancement and denoising method via sequential decomposition. In: IEEE International Symposium on Circuits and Systems (ISCAS) (2018)
61. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention (MICCAI) (2015)
62. Sakaridis, C., Dai, D., Van Gool, L.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: International Conference on Computer Vision (ICCV) (2019)
63. Sakaridis, C., Dai, D., Hecker, S., Van Gool, L.: Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In: European Conference on Computer Vision (ECCV) (2018)
64. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision (IJCV)* **126**(9), 973–992 (2018)
65. Salmon, J., Harmann, Z., Deledalle, C.A., Willett, R.: Poisson noise reduction with non-local pca. *Journal of mathematical imaging and vision* **48**(2), 279–294 (2014)
66. Sharma, V., Diba, A., Neven, D., Brown, M.S., Van Gool, L., Stiefelhagen, R.: Classification-driven dynamic image enhancement. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4033–4041 (2018)
67. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. *IEEE Transactions on image processing (TIP)* **15**(2), 430–444 (2006)
68. Shen, L., Yue, Z., Feng, F., Chen, Q., Liu, S., Ma, J.: Msr-net: Low-light image enhancement using deep convolutional network. arXiv preprint arXiv:1711.02488 (2017)
69. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *Computer Science* (2014)
70. de Stoutz, E., Ignatov, A., Kobyshev, N., Timofte, R., Van Gool, L.: Fast perceptual image enhancement. In: European Conference on Computer Vision Workshop (ECCVW), pp. 260–275 (2018)
71. Tao, L., Zhu, C., Song, J., Lu, T., Jia, H., Xie, X.: Low-light image enhancement using cnn and bright channel prior. In: IEEE International Conference on Image Processing (ICIP) (2017)
72. Tao, L., Zhu, C., Xiang, G., Li, Y., Jia, H., Xie, X.: Llcnn: A convolutional neural network for low-light image enhancement. In: Visual Communications and Image Processing (VCIP) (2017)
73. Wang, R., Zhang, Q., Fu, C.W., Shen, X., Zheng, W.S., Jia, J.: Underexposed photo enhancement using deep illumination estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
74. Wang, S., Zheng, J., Hu, H.M., Li, B.: Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing (TIP)* **22**(9), 3538–3548 (2013)
75. Wang, W., Wei, C., Yang, W., Liu, J.: Gladnet: Low-light enhancement network with global awareness. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG) (2018)
76. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing (TIP)* **13**(4), 600–612 (2004)
77. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. *British Machine Vision Conference (BMVC)* (2018)
78. Xu, J., Zhang, L., Zhang, D.: A trilateral weighted sparse coding scheme for real-world image denoising. European Conference on Computer Vision (ECCV) (2018)
79. Xu, L., Lu, C., Xu, Y., Jia, J.: Image smoothing via l<sub>0</sub> gradient minimization. In: ACM Transactions on Graphics (TOG), vol. 30, pp. 174–185 (2011)
80. Yamashita, H., Sugimura, D., Hamamoto, T.: Low-light color image enhancement via iterative noise reduction using rgb/nir sensor. *Journal of Electronic Imaging* **26**(4), 043017 (2017)
81. Yeganeh, H., Wang, Z.: Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing (TIP)* **22**(2), 657–667 (2013)
82. Ying, Z., Li, G., Gao, W.: A bio-inspired multi-exposure fusion framework for low-light image enhancement. arXiv preprint arXiv:1711.00591 (2017)
83. Ying, Z., Li, G., Ren, Y., Wang, R., Wang, W.: A new low-light image enhancement algorithm using camera response model. In: IEEE International Conference on Computer Vision (ICCV) (2017)
84. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing (TIP)* **26**(7), 3142–3155 (2017)

- 
- 85. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European Conference on Computer Vision (ECCV) (2016)
  - 86. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE conference on computer vision and pattern recognition (CVPR) (2018)
  - 87. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: European conference on computer vision (ECCV) (2018)
  - 88. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. arXiv preprint arXiv:1905.04161 (2019)