Low-Light Image Enhancement via a Deep Hybrid Network

Wenqi Ren, *Member*, IEEE, Sifei Liu, Lin Ma, Qianqian Xu, Xiangyu Xu, Xiaochun Cao, *Senior Member*, IEEE, Junping Du, and Ming-Hsuan Yang, *Fellow*, IEEE

***Abstract*—Camera sensors often fail to capture clear images or videos in a poorly-lit environment. In this paper, we propose a trainable hybrid network to enhance the visibility of such degraded images. The proposed network consists of two distinct streams to simultaneously learn the global content and salient structures of the clear image in a unified network. More specifically, the content stream estimates the global content of the low-light input through an encoder-decoder network. However, the encoder in the content stream tends to lose some structure details. To remedy this, we propose a novel spatially variant recurrent neural network (RNN) as an edge stream to model edge details, with the guidance of another auto-encoder. Experimental results show that the proposed network performs favorably against the state-of-the-art low-light image enhancement algorithms.**

***Index Terms*—Low-light image enhancement, convolutional neural network, recurrent neural network.**

I. Introduction

Images captured in the poorly light environment are often of low visibility and affect many high-level computer vision tasks such as detection and recognition. These all make low-light image enhancement a highly desirable technique. The formation of a low-light image can be modeled as [1]

$$L(x) = R(x) \circ T(x), \tag{1}$$

where $L(x)$ and $R(x)$ are the degraded image and its desired recovery, respectively. Furthermore, $T(x)$ represents the illumination map, *i.e.*, the light intensity on the scene in the image, and the operator $\circ$ means element-wise multiplication.

Since estimating the clear image $R(x)$ and the illumination $L(x)$ from a single observed image is an ill-posed inverse problem, numerous low-light image enhancement methods have been proposed [5]–[8] in recent years. Early approaches [9]–[13] focus on contrast enhancement to recover the visibility of dark regions. However, these global enhancement based methods may result in detail loss in some local areas because a global processing cannot ensure all local areas be well enhanced. Another line of research [2] tries to recover a sharp image by decomposing an input into the reflectance and the illumination components based on Retinex theory [14]. These methods either treat the reflectance as the final enhanced result or generate the result by manipulating the illumination. However, those methods may suffer from over- or under-enhancement (Figure 1(c) and (d)). The main reason is that these approaches using a physical model without considering the camera response characteristics [15]. Moreover, the errors produced by the estimation of the two factors can be accumulated when combined together.

Recently, deep convolutional neural networks (CNNs) have been applied to image enhancement [3], [4], [16]–[24]. Ignatov et al. [4] learn a mapping between photos from mobile devices and a DSLR camera based on an end-to-end residual network. This model uses a perceptual error function that combines content, color, and adversarial losses. We note that edge details are critical in image enhancement. However, this method does not particularly consider edge information when enhancing degraded inputs. In addition, bilateral grid processing is embedded in a neural network by Gharbi et al. [3] for real-time image enhancement. However, the method requires producing affine coefficients before obtaining outputs, which lacks direct supervision from the targets. For computer vision tasks, the number of affine coefficients is usually very large, which becomes the performance and speed bottlenecks [25].

To preserve the image naturalness and generate more accurate enhancement results, we present an automatic low-light image enhancement method based on a hybrid neural network. We achieve this through four key strategies: 1) We brighten the input images by a *content stream* that consists of an encoder and a decoder, where the encoder is used to capture the context of the low-light input and the decoder is used to estimate most of the scene using the learned representations from the encoder. 2) Since the encoder network in the content stream loses some image details, we propose an *edge stream* network, by combining a spatially variant RNN, to incorporate edge-aware feature maps and predict accurate image structures. The spatially RNNs [26] have been used for many low-level tasks. However, [26] models a unit response of the RNN filter and cannot model signals with notable energy variations (*e.g.*, from

This work is supported by the National Natural Science Foundation of China (No. U1736219, U1605252, U1803264, 61532006, 61772083, 61802403), National Key R&D Program of China (Grant No. 2018YFB0803701), Beijing Natural Science Foundation (No. L182057), and CCF-Tencent Open Fund. (Corresponding author: Xiaochun Cao)

W. Ren and X. Cao are with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (e-mail: rwq.renwenqi@gmail.com; caoxiaochun@iie.ac.cn).

S. Liu is with the NVIDIA Learning and Perception, Santa Clara, CA 95051, USA (e-mail: sifeil@nvidia.com)

L. Ma is with the Tencent AI Lab, Shenzhen 518000, China (e-mail: forest.linma@gmail.com).

Q. Xu is with the Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China (e-mail: xuqianqian@ict.ac.cn).

X. Xu is with SenseTime Research, Beijing 100084, China (e-mail: xuxiangyu2014@gmail.com).

J. Du is with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: junpingd@bupt.edu.cn).

M.-H. Yang is with the School of Engineering, University of California at Merced, Merced, CA 95343 USA (e-mail: mhyang@ucmerced.edu).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

(a) Ground truth MSE$_{hsv}$ | (b) Input 2354.8 | (c) WVM [2] 806.27 | (d) LIME [1] 477.24 | (e) HDRNet [3] 252.65 | (f) DSLR [4] 818.01 | (g) Our result **63.21**

Fig. 1. Enhanced results on a low-light image. The results generated by WVA [2] and DSLR-Q [4] tend to be dark throughout the region. Although HDRNet [3] yields a bright result, the color is too dazzling than it should be. In contrast, our method learns to reconstruct realistic results with clear structures, the color of the enhanced image is closest to the ground truth. We show the color constancy MSE$_{hsv}$ of different methods by computing MSE in the HSV space.

a low-light image to a day-light output) since the sum of the filter coefficients is equal to 1. In contrast, we improve the spatially variant linear RNNs by introducing two independent weight maps for the hidden state and image pixels. The weight maps associated with the output from the *content stream* reveal important image structures. 3) Since increasing the contrast of the details may make noise more visible, we mitigate this issue by adding a small level of Gaussian noise in our training data. Therefore, our proposed model can suppress noise to some extent. 4) We further incorporate perceptual and adversarial losses [16] to improve the visual quality of the enhanced results. Taking these four strategies together allows us to perform effective low-light image enhancement.

The contributions of this work are as follows.

1) We present a hybrid neural network in which the *content stream* is used to predict the scene information of the input, and the *edge stream* is devoted to edge details learning. This designed network is able to recover more accurate scene content.
2) We propose a spatially variant RNN by introducing two independent weight maps as input features and hidden states of a RNN. The RNNs model the internal structure of an image, such as edges, which plays an important role in low-light image enhancement.
3) We train the hybrid network with the perceptual and adversarial loss, producing visually pleasing enhanced images that perform favorably against state-of-the-art methods.

II. Related Work

There exist three main approaches for low-light image enhancement: Histogram Equalization (HE) based methods, algorithms that using Retinex theory, and data-driven approaches.

Histogram equalization. HE is a widely-used technique for lightening images with low visibility due to its simplicity. HE methods can be classified into two categories: global and local based approaches. Global based HE methods [27]–[30] use the histogram information of the entire input image for the transformation function estimation. Though these global approaches are suitable for overall enhancement, some high-frequency gray levels in the input tend to dominate the other low-frequency gray levels. To overcome this problem, local based HE approaches [10], [31], [32] use a sliding window over the image and only take pixels inside this window into account. However, Local HE requires high computational cost and sometimes causes over-enhancement in some portion of the image.

Retinex theory. Since HE based methods lack physical explanation, a physically meaningful model [2], [33] is proposed for image enhancement. Retinex theory [34] assumes that the amount of light reaching observers can be decomposed into illumination and scene reflection [35]–[37]. Wang et al. [38] propose a naturalness preserved enhancement algorithm for non-uniform illumination images. In [2], a weighted variational model is proposed to adjust the illumination by fusing multiple derivations of the initially estimated illumination map. However, this algorithm tends to lose the realism of regions with rich textures due to the blindness of illumination structure. Guo et al. [1] first estimate the illumination map by finding the maximum intensity of each pixel in RGB channels then exploit the structure of the illumination to refine the output.

Although Retinex based methods could efficiently emphasize the details of images, they may destroy the naturalness of images. In addition, according to the Retinex theory, it is unreasonable to treat only the reflectance layer as the enhanced image [39].

Data-driven approaches. Kang et al. [6] propose an exemplar-based method which uses a distance metric learning technique to determine a similarity measure that would map all images with similar enhancement requirements in the similar regions in image space. Caicedo et al. [40] further exploit the personalization aspect of the method and use collaborative filtering to discover clusters of user preferences. While the methods in [6] and [40] are non-parametric, Bychkovsky et al. [41] take a parametric approach to reproduce the global tonal remapping function by training on input-output image pairs. However, this remapping function is only learned for the luminance channel and is tested on a dataset where the amount of white balance is known, limiting its practicality.

Recently, Ignatov et al. [4], [42] present an end-to-end deep learning approach to translate ordinary photos taken by low-end cameras into DSLR-quality images. However, these methods focus on some specific devices, *e.g.*, *iPhone 3GS*, *BlackBerry Passport*, and *Sony Xperia Z*. To accelerate a wide variety of image processing operators, Chen et al. [43] use a fully-convolutional network to model the operators. In [44], a dataset of raw short-exposure low-light images is introduced.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

Then the authors train a fully convolutional network to avoid the noise amplification and error accumulation and demonstrate promising results on the raw dataset.

In contrast to previous works, we propose a novel hybrid network consisting of content and edge streams for general low-light image enhancement. Our method uses a spatially variant RNN to determine local structure information and a residual encoder-decoder to predict the main content of the output.

III. Our Method

We propose a novel hybrid network architecture to perform low-light image enhancement. We perform most of the inference on an input in the RGB space using the content stream (Figure 2, top). In addition, the salient edge stream (Figure 2, bottom) works at both RGB and gradient space. This stream has the critical role of capturing high-frequency effects and preserving salient structures. For this purpose, we introduce a novel spatially variant RNN inspired by [26]. Our architecture then fuses these two paths to yield the final result.

A. Content Stream

Inspired by the effectiveness of encoder-decoder networks in image denoising [45], dehazing [46], [47], inpainting [48], matting [49], and harmonization [50], we construct our content branch based on the residual encoder-decoder architecture with specific designs for content prediction as follows: 1) the first two convolutional layers in the encoder are changed to dilated convolution [51] to enlarge the receptive field, and 2) features of deconvolution modules in edge stream are concatenated together to obtain more details during the up-sampling stage.

The residual encoder-decoder network has three convolution modules, and each consists of several convolutional layers, ReLU and skip links. Specifically, the features from the 1st, 2nd, and 3rd convolutions are with sizes of 1/2, 1/4, 1/8 of the input image size, respectively. The corresponding decoder introduces up-sampling operation to enlarge the feature maps (upper branch in Figure 2).

B. Edge Stream

Besides the content prediction, we introduce an edge stream (Figure 3) with spatially variant RNNs to learn the weight maps conditioned on the input image and the corresponding gradients. RNNs [26], [52] have been developed for low-level vision tasks. For example, a spatially RNN is proposed in [26], which transfers the previous hidden state $h\left[k-1\right]$ to the current state $h\left[k\right]$ with the input image pixel $x\left[k\right]$ at location k. Specifically, the spatial recurrent relation in one-dimensional (1D) can be modeled by

$$h\left[k\right]=\left(1-p\left[k\right]\right)\circ x\left[k\right]+p\left[k\right]\circ h\left[k-1\right], \tag{2}$$

where $p\left[k\right]$ is the weighting factor that balances the contributions between $x\left[k\right]$ and $h\left[k\right]$. Deep CNN relies on the image content are used to learn the corresponding weight map p. However, Eq. 2 is a normalized filter which has a unit gain at some specified frequency. For example, a low-pass filter commonly has unit gain, which implies that its discrete impulse response should sum to one. As such, this method cannot be directly applied on a low-light image enhancement task, since the overall energies of the low-light and day-light images are of significant difference. As shown in Figure 4, the output in (b) is still dark when directly use the normalized RNNs.

Inspired by the spatially variant RNN in Eq. 2, we propose an edge stream to remedy structure information losses of the low-light images. More specifically, we propose an improved spatially variant RNN model:

$$h\left[k\right]=g\left[k\right]\circ x\left[k\right]+p\left[k\right]\circ h\left[k-1\right], \tag{3}$$

Different from Eq. 2, a new weight map g is introduced, which is independent with p. As such, current image pixel information $x\left[k\right]$ and the previous hidden state $h\left[k-1\right]$ can be more flexible fused. For example, when $p\left[k\right]$ is close to zero, it cuts the propagations from $h\left[k-1\right]$ to $h\left[k\right]$. Therefore, only the current image pixel information is considered, with the edge information preserved. On the other hand, $p\left[k\right]$ with a larger value maintains the propagation from $h\left[k-1\right]$ to $h\left[k\right]$. As such, the edge feature tends to be smoothed at non-edge locations.

Figure 3 illustrates one single directional RNN (left-to-right) example. The input information is composed of the low-light image and its corresponding gradients, as illustrated in Figure 2. The deep CNNs rely on the input information to generate the weight maps g and p. Multi-scale features x are further composed from the input image. One directional spatially variant RNN, specifically left-to-right (red arrows in Figure 3), takes the generated weight maps g and p as well as the multi-scale features x, to recurrent generate the hidden state, namely the edge features. Note that for each specific direction, the input image is considered as a group of 1D sequences. In addition to left-to-right, we consider the other three directions: right-to-left, top-to-bottom, and bottom-to-top. For left-to-right and right-to-left, each row of the input image is regarded as one sequence. If one input pixel is regarded as $x[k]$, its left pixel is $x[k-1]$ for the left-to-right direction, while its right pixel is $x[k-1]$ for the right-to-left case. Similarly, each column is regarded as one sequence for top-to-bottom and bottom-to-top directions.

In this paper, we also adopt an encoder-decoder architecture with additional skip links to compute the pixel-wise weights $g[k]$ and $p[k]$. We find that this network shares similar properties with the content branch and their feature representations are in similar scales, which enables plausible connections to the content stream as illustrated in Figure 2. Additionally, to generate the multi-scale features x, we use down-convolutional layers of the input image with ratios of $\{1/2, 1/4, 1/8\}$, then resize them to the original size and concatenate them together with the original image. Therefore, $x[k]$ in our spatially variant RNNs can reach to a more global range via processing on coarse scales. With the multi-scale features x and the learned weight maps g and p, our RNNs rely on Eq. 3 to scan the image from four different directions. Therefore, four hidden activation maps are generated to learn different edge relevant features. We integrate these features by selecting the optimal

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

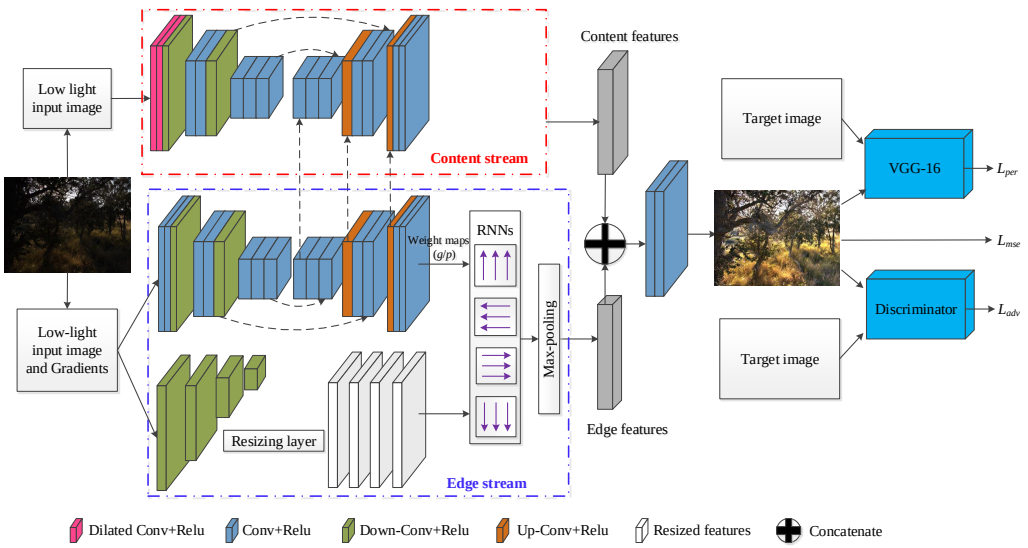


Fig. 2. Proposed low-light image enhancement architecture. Our model consists of two streams, the top content stream is a residual encoder-decoder which aims to restore most of the scene. The bottom edge stream focuses on the salient edge prediction via spatially variant RNNs. To construct communications between two streams, we bridge two networks during the up-sampling stage. In addition to the MSE loss function, we also adopt the perceptual and adversarial losses to further improve the visual quality.

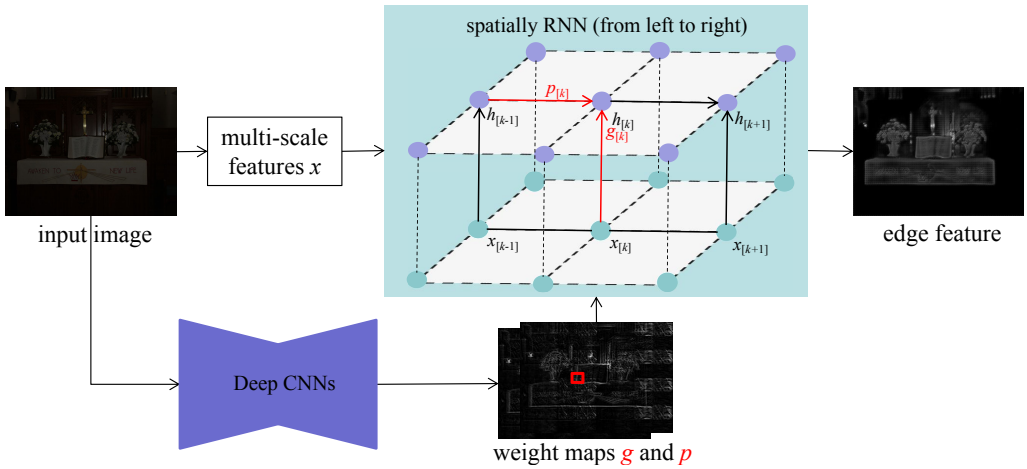


Fig. 3. An example of learning edge information by the spatial RNN from different directions in the proposed *edge stream*. Here we only show the RNN from left to right. The deep CNN generates weight maps that guide the propagation of the RNNs.

direction based on the maximum response at each location. This is carried out by a node-wise max pooling, which effectively selects the maximally responded direction as the desired salient edge information to be propagated.

After the content and edge-relevant features are learned by our hybrid network, we fuse these features followed by two additional convolutional layers as shown in Figure 2.

C. Loss Function

We build our loss function under the assumption that the overall perceptual image quality can be improved by optimizing three independent losses: i) MSE loss, ii) adversarial loss and iii) perceptual loss. We now define loss functions for each component.

MSE loss. To measure the content difference between the enhanced and ground truth image, we use the Euclidean distance between the obtained results and target images. Therefore, the loss of the generator network G can be written as:

$$\mathcal{L}_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N}\left\|G(R) - I\right\|^2, \tag{4}$$

where N denotes the number of the images in each process.

Perceptual loss. To improve the enhancement effect of the generator, we also use the perceptual loss to constrain the generator. The effect of the perceptual loss has been demonstrated in image restoration and other relative fields [53], [54]. The perceptual loss is defined as the Euclidean distance between the feature representations of a reconstructed image $G(R)$ and the reference image I from the ReLU activation layers of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

(a) (b) (c)

Fig. 4. Since the spatially RNN [26] is equivalent to a normalized filter, the output of RNNs is still a dark image (b) given a low-light input (a). In contrast, our improved spatially variant RNN learns salient edge relevant features as shown in (c).

pre-trained 16 layer VGG network [55].

$$\mathcal{L}_{\text{per}} = \frac{1}{C_j W_j H_j} \left\| \phi_j\{G(R)\} - \phi_j\{I\} \right\|^2, \tag{5}$$

where ϕ_j indicates the feature map obtained by the j-th convolution by the VGG16 network. In addition, C_j, W_j, and H_j denote the number, height, and width of the feature maps.

TABLE I
MODEL PARAMETERS OF THE DISCRIMINATOR. EVERY CONVOLUTION LAYERS ARE ACTIVATED WITH A LEAKYRELU LAYER.

Layer	Kernel dimension	Stride	Output size
Input	-	-	128×128
Conv1-1	$32 \times 5 \times 5$	2	64×64
Conv1-2	$64 \times 3 \times 3$	1	64×64
Conv2-1	$64 \times 5 \times 5$	2	32×32
Conv2-2	$128 \times 3 \times 3$	1	32×32
Conv3-1	$128 \times 5 \times 5$	2	16×16
Conv3-2	$256 \times 3 \times 3$	1	16×16
Conv4-1	$256 \times 5 \times 5$	2	8×8
Conv4-2	$512 \times 3 \times 3$	1	8×8
Conv5-1	$512 \times 5 \times 5$	2	4×4
Conv5-2	$1 \times 4 \times 4$	1	1×1
Sigmoid	-	-	-

Adversarial loss. Recently, the adversarial loss has been used extensively in low-level vision tasks. To promote the performance to be realistic, on the account of learning a generative adversarial network via distinguishing the real image and the fake image to simulate the restored images, we hope the adversarial model can assist to train the generator G that generates enhanced images which can cheat the discriminator D to distinguish from real images. The adversarial loss is defined as follows.

$$\begin{aligned} \mathcal{L}_{adv} = & \mathop{\mathbb{E}}_{I \backsim p_{\text{clear}}(I)} \left[\log D(I) \right] \\ & + \mathop{\mathbb{E}}_{R \backsim p_{\text{low}}(L)} \left[\log \left(1 - D(G(L)) \right) \right]. \end{aligned} \tag{6}$$

We build the discriminator as shown in Table I. The discriminator D consists of ten convolutional layers each followed by a LeakyReLU nonlinearity. The kernel size of the first, third, fifth, seventh and ninth convolutional layers is 5×5 and the stride step size is 2. Other convolutional layers are with the size of 3×3 and stride of 1. A sigmoidal activation function is applied to the outputs of the last convolutional layer and produces a probability that the input image is the same as ground truth.

Fig. 5. Examples of our training data. The first row is the captured real low-light photographs. The second row shows the retouched output by professional retoucher.

Total loss. By combining the MSE loss, perceptual loss and adversarial loss, our final loss function is defined as a weighted sum of these losses

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mse}} + \lambda_p \mathcal{L}_{\text{per}} + \lambda_a \mathcal{L}_{\text{adv}}, \tag{7}$$

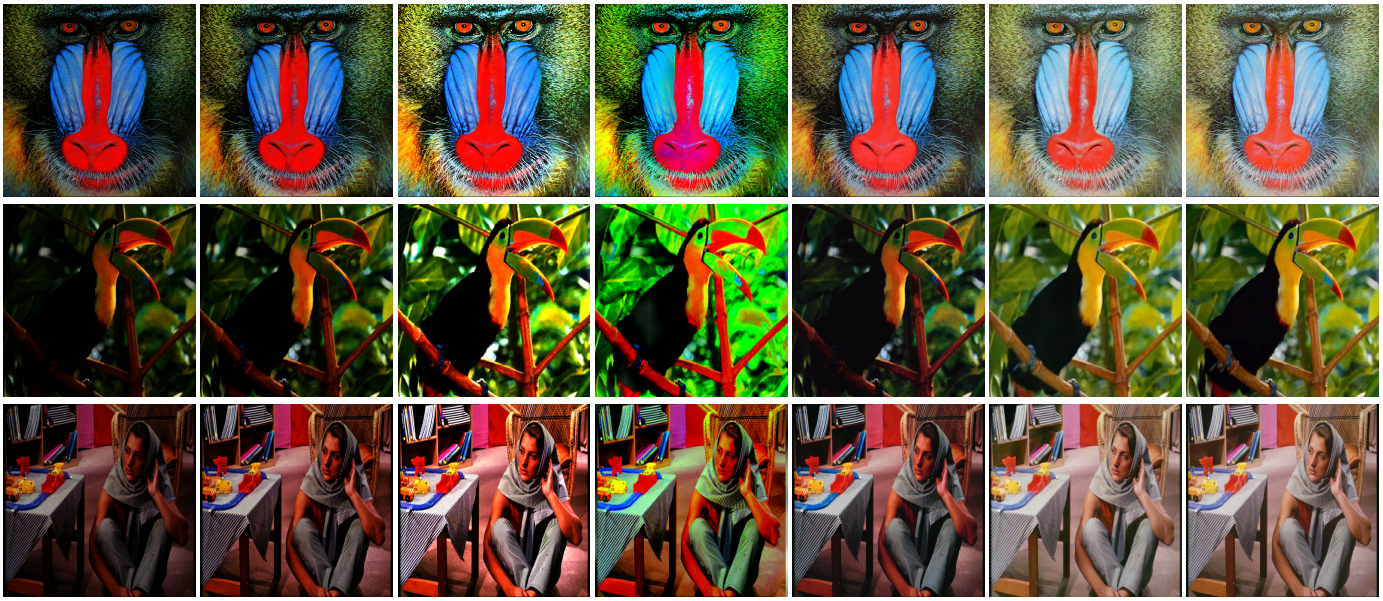
where λ_p and λ_a are the balanced weights. We analyze the roles of different losses in Section IV-B.

D. Implementation Details

During training, we use a batch size of 10, and patch size of 128×128. The balanced weights in Eq. 7 are set to $\lambda_p = 0.05$ and $\lambda_d = 1e^{-3}$. We illustrate our hybrid network architecture in Figure 2, our low-light image enhancement network is trained using an end-to-end style. There are 3 convolution layers in each convolution block in our residual encoder-decoder network in both content and edge streams. We employ the dilated convolutions in the first two convolutional layers in the content stream with a dilated factor of 2. As a result, the latent contents can be extracted by this network with a larger receptive field. We use ADAM [56] optimizer for training. The initial learning rate is 0.0001 and decreased by 0.5 every 8,000 iterations. For all the results reported in the paper, we train the network for 20,000 iterations. Default values of β_1 and β_2 are used, which are 0.9 and 0.999, respectively, and we set weight decay to 0.00001.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

Fig. 6. The 10 testing images.



(a) Inputs (b) WVM [2] (c) LIME [1] (d) HDRNet [3] (e) DSLR-Q [4] (f) Ours (g) Ground truth

Fig. 7. Enhanced results on synthetic images where the ground truths are shown in (g). The results by WVM [2] and DSLR-Q [4] still appear dark. Although LIME [1] and HDRNet [3] generate bright images as shown in (c) and (d), these results have some color distortions. In contrast, our results are close to the ground truths.

IV. Experimental Results

We compare the proposed algorithm with several state-of-the-art methods in terms of pixel fidelity, color constancy, and visual effect. The implementation code, as well as the dataset, will be made available to the public.

A. Training Data

Generating realistic training data is a major challenge for low-light image enhancement task since ground truth data cannot be easily collected. To overcome this problem, Lore et al. [8] generate low-light images by darkening natural images using nonlinear Gamma correction and use the original images as the ground truths. However, Gamma correction is a global adjustment for the whole image and cannot take into account the image content. In [4], Ignatov et al. collect a large-scale dataset of over $6,000$ photos taken synchronously by a DSLR camera and 3 low-end cameras of smartphones in a wide variety of conditions. But these captured image pairs are not perfectly aligned and need additional nonlinear transformation.

In contrast, our training data is carefully selected from the MIT-Adobe FiveK dataset [41]. This dataset includes 5000 captured photographs, and each of them is retouched by 5 trained photographers with different styles including sunset mood and day like look. Therefore, this dataset includes 5 sets of $5,000$ input and output pairs that enable supervised learning. Since we focus on low-light image enhancement, we carefully select 336 input and output pairs that meet our requirements from the Adobe FiveK dataset [41]. The selected photographs cover a broad range of scenes and subjects. We show some examples in Figure 5. Due to the relatively low number of training images, we further augment the dataset by using rotation, flip, clipping, noise, and a small degree of gamma correction (i.e., $\gamma \in (2, 4)$) to further darken images.

B. Quantitative Results

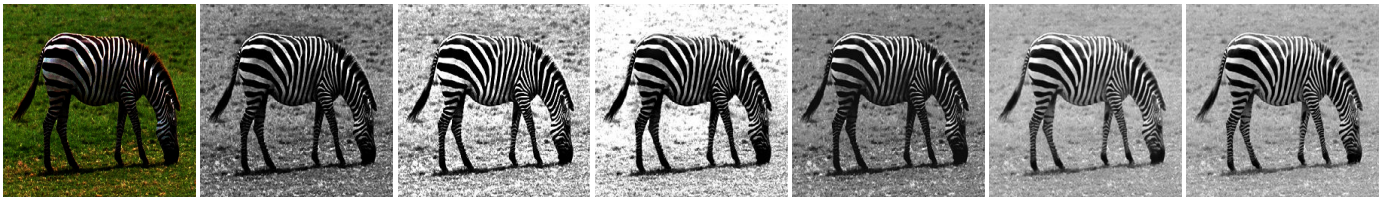
In this section, we quantitatively evaluate the state-of-the-art image enhancement methods as well as the proposed algorithm on two datasets.

Artificially darkened images. We first evaluate the performance of our algorithm against the state-of-the-arts including Retinex-based methods WVM [2] and LIME [1], and CNN-based approaches HDRNet [3] and DSLR-Q [4] on 10 widely used testing images shown in Figure 5. Testing images are darkened by gamma correction with $\gamma = 3$ In addition, 1% Gaussian noise is added to each image. We use four examples: *Baboon*, *Bird*, *Barbara*, and *Lena* for illustration. Figure 7(a) shows the input darkened images which are synthesized from the sharp images in (g). As the method of WVM [2] loses the realism of regions with rich textures due to the blindness of illumination structures, this approach tends to result in darker

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

TABLE II
AVERAGE PSNR AND SSIM OF ENHANCED RESULTS ON THE SYNTHETIC IMAGES.

	PSNR/SSIM							
	WVM [2]	LIME [1]	HDRNet [3]	DSLR-Q [4]	$\mathcal{L}_{\text{mse}}$	$\mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{per}}$	$\mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{adv}}$	Ours $\mathcal{L}_{\text{total}}$
airplane	15.10/0.77	17.65/0.73	17.27/0.31	16.87/0.75	**30.72**/0.98	29.66/**0.99**	29.96/**0.99**	30.64/0.98
baboon	12.00/0.66	14.62/0.73	15.15/0.55	15.03/0.77	27.14/**0.95**	**27.42/0.95**	27.18/**0.95**	27.06/**0.95**
baby	12.25/0.54	14.55/0.68	14.62/0.49	12.60/0.62	28.79/0.95	**30.39/0.96**	30.21/0.95	30.06/**0.96**
bird	14.15/0.37	17.24/0.63	16.250.51	15.34/0.52	**28.36/0.93**	26.46/0.91	25.82/0.90	28.16/**0.93**
butterfly1	12.96/0.60	15.01/0.70	15.80/0.63	15.06/0.72	28.72/0.97	29.52/0.97	**30.37/0.98**	29.96/0.97
barbara	13.10/0.60	15.60/0.64	16.76/0.54	15.89/0.72	25.98/0.97	**28.10/0.98**	27.83/0.97	27.10/**0.97**
lena	11.83/0.62	13.59/0.71	16.32/0.70	13.30/0.67	27.69/0.96	30.44/**0.97**	**30.52**/0.96	29.80/**0.97**
butterfly2	13.11/0.64	15.06/0.70	17.11/0.65	14.57/0.66	25.16/0.97	**27.75/0.97**	27.04/0.97	26.85/**0.97**
pepper	13.67/0.68	15.61/0.74	14.27/0.60	16.85/0.78	24.93/**0.95**	25.86/0.94	25.95/0.94	**26.50/0.95**
zebra	12.56/0.65	15.80/0.73	13.94/0.48	14.59/0.77	27.88/0.95	**29.04/0.96**	28.56/0.95	28.14/**0.96**
Average	13.07/0.61	15.47/0.70	15.75/0.54	15.01/0.70	27.54/0.96	**28.46/0.96**	28.34/**0.96**	28.43/**0.96**

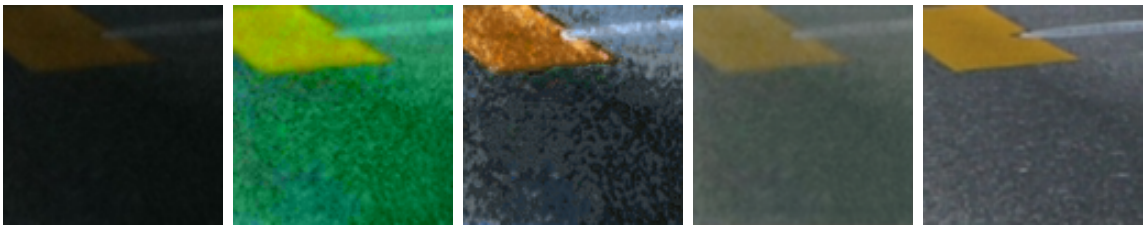


(a) Inputs (b) WVM [2] (c) LIME [1] (d) HDRNet [3] (e) DSLR-Q [4] (f) Ours (g) Ground truth

Fig. 8. Comparison of enhanced estimation in the HSV domain. The results look gray because we only show the estimated V-channel without Hue and Saturation. Our results on the V channel is closest to the ground image.

TABLE III
AVERAGE MSE OF ENHANCED RESULTS ON HSV SPACE OF 10 SYNTHETIC IMAGES.

Metrics	NPE [38]	WVM [2]	LIME [1]	HDRNet [3]	DSLR-Q [4]	Ours
MSE_{hsv}	606.87	873.69	441.68	618.13	647.49	**45.18**
MSE_{h}	0.0304	**0.0290**	0.0327	0.0812	0.1421	0.0568
MSE_{s}	0.1115	0.1101	0.1071	0.2208	0.0749	**0.0136**
MSE_{v}	1820.5	2620.9	1324.9	1854.1	1942.3	**135.47**



(a) Inputs (b) HDRNet [3] (c) DSLR-Q [4] (d) Ours (e) Ground truth

Fig. 9. Visual comparison of results on the DPED dataset. The proposed method performs favorably against the state-of-the-art approaches.

TABLE IV
AVERAGE PSNR AND SSIM OF ENHANCED RESULTS ON THE DPED DATASET [4].

	PSNR/SSIM						
	WVM [2]	HDRNet [3]	PLSR [53]	DSLR-Q [4]	Ours w/o $\mathcal{L}_{per}$	Ours w/o $\mathcal{L}_{adv}$	Ours
iphone	**22.22**/0.7320	18.31/0.6432	20.32/0.9161	20.08/0.9201	20.70/**0.9236**	21.51/0.9218	20.90/0.9129
blackberry	20.87/0.7311	18.08/0.6366	20.11/0.9298	20.07/0.9328	20.96/0.9324	**21.15**/0.9210	20.62/**0.9331**
sony	19.74/0.7250	18.06/0.7895	21.33/0.9434	21.81/**0.9437**	22.78/0.9351	**23.46**/0.9293	22.32/0.9375

results as shown in Figure 7(b). Although the reconstructed results by LIME [1] are better than those by [2], the enhanced images generated by LIME [1] tend to have some color distortions. For example, the colors of the feathers are changed from black to purple in the fourth row in Figure 7(c). We note that HDRNet [3] produces slightly color distortions as shown in Figure 7(d). The main reason is that HDRNet expects as input a Bayer RAW RGB image[1], while our testing data are sRGB (standard Red Green Blue) images. The recovered results by DSLR-Q method [4] are better than those by [1]–[3], but the colors are still darker than the ground truth. In contrast, the enhanced results by the proposed algorithm in Figure 7(f) are close to the ground truth sharp images. Table II shows that the proposed algorithm performs favorably on each image against the state-of-the-art enhancement methods [1], [2], [4], [38] in terms of PSNR and SSIM. The improvement is significant as the proposed algorithm achieves better results by up to 13 dB in terms of average PSNR, and up to 0.25 in terms of average SSIM metric on the 10 darkened images as shown in the last row in Table II.

Color constancy. In addition to enhancing the low-light image, the proposed model also recovers authentic color even though we minimize on RGB space. Figure 8 provides an example. The V channels recovered by WVM [2] and DSLR-Q [4] are darker than the ground truth, while the results by LIME [1] and HDRNet [3] are brighter than the ground truth in (g). In contrast, our enhanced image is much more similar to the ground truth. To evaluate the performance of the different algorithms, we compute the MSE error between the ground truths and recovered results in HSV domain on the 10 testing images. As shown in Table III, our algorithm performs favorably against the state-of-the-art methods on all the three H, S, and V channels. Since H and S channels are expressed in a range from just 0 to 1, while the intensity of Value channel between 0 and 255. The error in V channel (MSE_v) dominates the error in HSV (MSE_{hsv}) color space, which demonstrates that our algorithm can recover the real color information from dark images.

DPED dataset. We then quantitatively compare the proposed model with state-of-the-art methods on the DPED dataset [4]. In total, this dataset includes 4353 patches from iPhone smartphone, 2436 from BlackBerry and 2521 patches from Sony, for each smartphone photo there is a corresponding photo from the Canon DSLR. Most of the images captured by smartphones have low brightness, and the images captured by Canon DSLR have a great visual effect and serve as ground truths.

We show two examples in Figure 9 to compare with the most related CNN based methods [3] and [4]. The results by HDRNet [3] tend to have some color distortion, while the results by DSLR-Q [4] still have some dark regions. Table IV displays the average PSNR and SSIM results on the DPED dataset. Since the proposed network is optimized from end to end under the MSE loss, it is not surprising to see its higher PSNR performance than other state-of-the-art methods. More appealing observation is that the proposed algorithm obtains better SSIM results against all competitors, even though SSIM is not directly referred to as an optimization criterion. Although the average PSNR achieved by the WVM approach [2] on *iPhone* dataset is higher than the results by other methods, its SSIM values are about 0.2 lower than our algorithm. In addition, as we employed a loss function that combines both the content loss (MSE), perceptual loss and the adversarial loss for training our network. We examine the effect of the perceptual and adversarial loss terms quantitatively. The PSNR and SSIM results are also shown in last three columns in Table IV. From these results, we observe that adding perceptual loss tends to increase PSNR, while the adversarial loss encourages to improve the SSIM value. We note that similar observations have been reported in [57], [58] that adversarial loss helps to improve visual effect, and SSIM measures beyond pixel-wise errors and is well-known to reflect the human perception more faithfully.

C. Qualitative Visual Results

We evaluate the proposed algorithm against the state-of-the-art enhancement methods [2]–[4] using five challenging real images in Figure 10. The lightness recovered by WVM [2] and DSLR-Q [4] is still somewhat dim and cannot effectively extract the information in dark regions as shown in Figure 10(b) and (d). The method of HDRNet [3] can augment the image details and enhance image visibility in (c). However, this method tends to over-enhance the colors in the generated images and results in some color distortions. For example, the color of the water is changed from gray to green in the first row in (c). In contrast, the recovered results by the proposed algorithm are visually more pleasing in dark regions without color distortions or artifacts.

D. Run Time

We select 40 images for all methods to run on the same machine. Table V summarizes the average run time of some representative methods on different image resolutions. WVM [2] and LIME [1] run on an Intel Core i7 CPU. HDRNet [3] and DSLR-Q [4] run on an Nvidia K80 GPU. In addition, we run our proposed model on both CPU and GPU for a fair comparison. The proposed method performs favorably against other state-of-the-art methods in terms of run time as shown in Table V.

TABLE V
AVERAGE TIME COST (SECONDS) WITH DIFFERENT IMAGE SIZES. WVM [2] AND LIME [1] RUN ON AN INTEL CORE I7 CPU AND HDRNET [3], DSLR-Q [4], AND OUR METHOD RUN ON A NVIDIA K80 GPU.

CPU/GPU	WVM	LIME	HDRNet	DSLR-Q	Ours
384×384	4.28/—	0.19/—	—/0.06	—/0.18	7.33/0.25
512×512	8.12/—	0.33/—	—/0.05	—/0.32	13.69/0.32

[1]https://github.com/mgharbi/hdrnet

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

(a) Inputs (b) WVM [2] (c) HDRNet [3] (d) DSLR-Q [4] (e) Ours

Fig. 10. Visual comparison for real image enhancement. The results generated by WVM [2] and DSLR-Q [4] still have some dark regions. HDRNet [3] suffer from over-enhancement as shown in (c). In contrast, the recovered image in (e) has rich details.

V. Analysis and Discussions

A. Comparison of Loss Functions

In this section, we compare the effects of different loss functions quantitatively and qualitatively. The PSNR and SSIM results by using $\mathcal{L}_{\text{mse}}$, $\mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{per}}$, $\mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{adv}}$, and $\mathcal{L}_{\text{total}}$ are shown in the last four columns in Table II, respectively. We observe that only using MSE loss achieves better results than the state-of-the-art methods [1], [2], [4], [38], but adding perceptual and adversarial losses could further improve the enhanced quality. Especially, the perceptual loss tends to increase PSNR, which achieves the highest PSNR value of 28.46 dB. While using the total loss function $\mathcal{L}_{\text{total}}$ tends to improve the SSIM value as shown in the last column in Table II.

Figure 11 shows two more visual comparisons between the results of our network trained with $\mathcal{L}_{\text{mse}}$, $\mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{per}}$, $\mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{adv}}$, and $\mathcal{L}_{\text{total}}$, respectively. As can be seen, all the results generated by our network have visually pleasant effect than the input in Figure 11(a). But adding the perceptual or adversarial losses could further invigorate the colors as shown in Figure 11(c)-(e).

B. Effectiveness of Edge Stream

In this section, we analyze how the edge stream helps improve recovering the edge in the results. The proposed edge stream is dedicated to learning the main structure of the input image, which greatly improves the final estimation of enhanced scenes. As stated in Section III-B, the features learned by edge stream mainly represent the salient edges. To better understand how the edge stream network affects our method, we conduct experimental results with different setups on the 10 synthetic images in Section IV-B. Table VI shows that using the edge stream performs better than only using the content stream by up to 0.7dB in terms of PSNR.

To further demonstrate the effectiveness of the edge stream, we show the visual results with and without using the edge stream in Figure 12. In addition, we also train the U-Net [59] using the same training data in this work to demonstrate the effectiveness of the edge stream. We note that the U-Net has a similar structure to our content branch but with a deeper architecture. Different from [59] which uses unpadded convolutions in the network, we employ padded convolutions to maintain the resolution of feature maps and outputs. As can be seen, both the recovered images in (b) and (c) have some blurry artifacts since the loss of information in the encoder process. In contrast, adding the edge stream could effectively recover the edge details as shown in Figure 12(d). These results indicate that the proposed edge stream plays a critical role in image enhancement.

C. Limitations

Although our results look more realistic, our generated results in Section IV-C are not as flamboyant as the method of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2019.2910412, IEEE Transactions on Image Processing

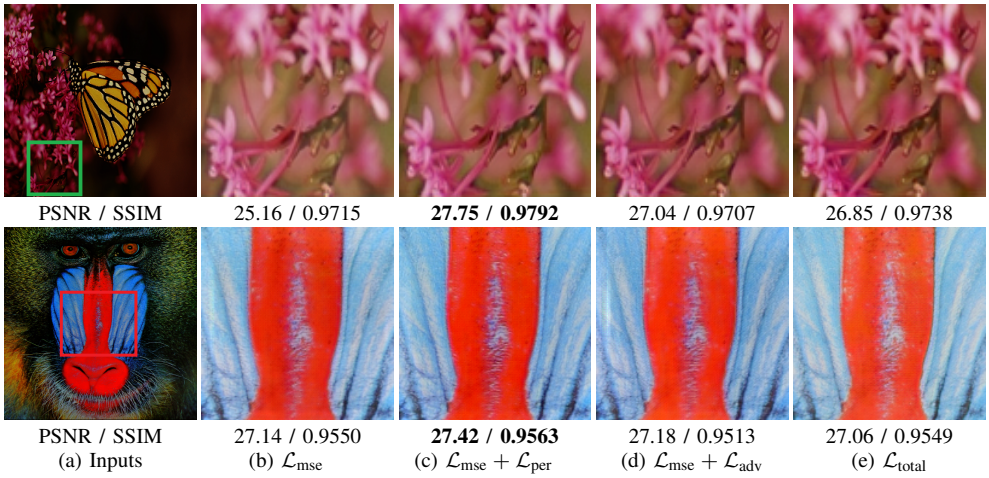


Fig. 11. Visual comparison of results from our model trained with different loss functions.

Fig. 12. Enhanced images with and without the proposed edge stream. Since U-Net [59] and the content stream loses some image details at the encoder process, the outputs in (b) and (c) have some blurry artifacts. However, the proposed edge stream can remedy the edge details effectively.

HDRNet [3]. Furthermore, we note that DSLR-Q [4] not only brighten the low-light input, but also output a high-resolution image. Future work will try to consider super-resolve the low-light input and find a trade-off between realistic and flamboyant colors.

TABLE VI
AVERAGE PSNR AND SSIM WITH AND WITHOUT EDGE STREAM.

	w/o edge stream	ours
PSNR/SSIM	27.67/0.9156	28.43/0.9662

VI. CONCLUSIONS

In this paper, we address the low-light image enhancement problem via a hybrid deep network. In the proposed deep model, the *content stream* is used to enhance the visibility of the low-light input and learn a holistic estimation of the scene content, while the *edge stream* network is devoted to refining the edge information using both input and its gradients based on an improved spatially variant RNN. In addition, we also combine the perceptual and adversarial loss functions with the proposed network to further improve the visual quality of the enhanced results. The quantitative and qualitative experimental results show that our method compares favorably against the state-of-the-art low-light image enhancement approaches.

REFERENCES

[1] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 982–993, 2017. 1, 2, 6, 7, 8, 9
[2] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, and X. Ding, "A weighted variational model for simultaneous reflectance and illumination estimation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2, 6, 7, 8, 9
[3] M. Gharbi, J. Chen, J. T. Barron, S. W. Hasinoff, and F. Durand, "Deep bilateral learning for real-time image enhancement," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 118, 2017. 1, 2, 6, 7, 8, 9, 10
[4] A. Ignatov, N. Kobyshev, K. Vanhoey, R. Timofte, and L. Van Gool, "DSLR-Quality photos on mobile devices with deep convolutional networks," *IEEE International Conference on Computer Vision*, 2017. 1, 2, 6, 7, 8, 9, 10
[5] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. 1
[6] S. B. Kang, A. Kapoor, and D. Lischinski, "Personalization of image enhancement," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1, 2
[7] S. J. Hwang, A. Kapoor, and S. B. Kang, "Context-based automatic local image enhancement," in *European Conference on Computer Vision*, 2012. 1
[8] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017. 1, 6
[9] H. Cheng and X. Shi, "A simple and effective histogram equalization approach to image enhancement," *Digital signal processing*, vol. 14, no. 2, pp. 158–170, 2004. 1
[10] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 2, 2007. 1, 2

[11] C. Wang and Z. Ye, "Brightness preserving histogram equalization with maximum entropy: a variational perspective," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 4, pp. 1326–1334, 2005. 1

[12] D. Sen and S. K. Pal, "Automatic exact histogram specification for contrast enhancement and visual system based quantitative evaluation," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1211–1220, 2011. 1

[13] T. Arici, S. Dikbas, and Y. Altunbasak, "A histogram modification framework and its application for image contrast enhancement," *IEEE Transactions on Image Processing*, vol. 18, no. 9, pp. 1921–1935, 2009. 1

[14] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977. 1

[15] Z. Ying, G. Li, Y. Ren, R. Wang, and W. Wang, "A new low-light image enhancement algorithm using camera response model," *IEEE International Conference on Computer Vision Workshop*, 2017. 1

[16] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu, "Automatic photo adjustment using deep neural networks," *ACM Transactions on Graphics*, vol. 35, no. 2, p. 11, 2016. 1, 2

[17] W. Ren, J. Zhang, L. Ma, J. Pan, X. Cao, W. Zuo, W. Liu, and M.-H. Yang, "Deep non-blind deconvolution via generalized low-rank approximation," in *Advances in Neural Information Processing Systems*, 2018, pp. 295–305. 1

[18] X. Xu, D. Sun, S. Liu, W. Ren, Y.-J. Zhang, M.-H. Yang, and J. Sun, "Rendering portraitures from monocular camera and beyond," in *European Conference on Computer Vision*, 2018, pp. 35–50. 1

[19] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European Conference on Computer Vision*, 2016, pp. 154–169. 1

[20] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "Msr-net: Low-light image enhancement using deep convolutional network," *arXiv preprint arXiv:1711.02488*, 2017. 1

[21] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2019. 1

[22] S. Li, I. B. Araujo, W. Ren, Z. Wang, E. K. Tokuda, R. H. Junior, R. Cesar-Junior, J. Zhang, X. Guo, and X. Cao, "Single image deraining: A comprehensive benchmark analysis," *arXiv preprint arXiv:1903.08558*, 2019. 1

[23] Y. Yuan, Y. Lyu, X. Shen, I. W. Tsang, and D.-Y. Yeung, "Marginalized average attentional network for weakly-supervised learning," in *International Conference on Learning Representations*, 2019. 1

[24] Y. Yuan, X. Liang, X. Wang, D.-Y. Yeung, and A. Gupta, "Temporal dynamic graph lstm for action-driven video object detection," in *IEEE International Conference on Computer Vision*, 2017. 1

[25] H. Wu, S. Zheng, J. Zhang, and K. Huang, "Fast end-to-end trainable guided filter," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[26] S. Liu, J. Pan, and M.-H. Yang, "Learning recursive filters for low-level vision via a hybrid neural network," in *European Conference on Computer Vision*, 2016. 1, 3, 5

[27] D. Coltuc, P. Bolon, and J.-M. Chassery, "Exact histogram specification," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1143–1152, 2006. 2

[28] H. Ibrahim and N. S. P. Kong, "Brightness preserving dynamic histogram equalization for image contrast enhancement," *IEEE Transactions on Consumer Electronics*, vol. 53, no. 4, 2007. 2

[29] G. Thomas, D. Flores-Tapia, and S. Pistorius, "Histogram specification: a fast and flexible method to process digital images," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 5, pp. 1565–1578, 2011. 2

[30] T. Celik and T. Tjahjadi, "Contextual and variational contrast enhancement," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3431–3441, 2011. 2

[31] J. A. Stark, "Adaptive image contrast enhancement using generalizations of histogram equalization," *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 889–896, 2000. 2

[32] C. Lee, C. Lee, and C.-S. Kim, "Contrast enhancement based on layered difference representation of 2d histograms," *IEEE Transactions on Image Processing*, vol. 22, no. 12, pp. 5372–5384, 2013. 2

[33] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal Processing*, vol. 129, pp. 82–96, 2016. 2

[34] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu, "Fast efficient algorithm for enhancement of low lighting video," in *IEEE International Conference on Multimedia and Expo*, 2011. 2

[35] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithms," in *IEEE International Conference on Computer Vision*, 2009, pp. 2335–2342. 2

[36] C. Rother, M. Kiefel, L. Zhang, B. Schölkopf, and P. V. Gehler, "Recovering intrinsic images with a global sparsity prior on reflectance," in *Advances in Neural Information Processing Systems*, 2011, pp. 765–773. 2

[37] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2752–2759. 2

[38] S. Wang, J. Zheng, H.-M. Hu, and B. Li, "Naturalness preserved enhancement algorithm for non-uniform illumination images," *TIP*, vol. 22, no. 9, pp. 3538–3548, 2013. 2, 7, 8, 9

[39] H. Yue, J. Yang, X. Sun, F. Wu, and C. Hou, "Contrast enhancement based on intrinsic image decomposition," *IEEE Transactions on Image Processing*, 2017. 2

[40] J. C. Caicedo, A. Kapoor, and S. B. Kang, "Collaborative personalization of image enhancement," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

[41] V. Bychkovsky, S. Paris, E. Chan, and F. Durand, "Learning photographic global tonal adjustment with a database of input/output image pairs," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2, 6

[42] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, "Wespe: Weakly supervised photo enhancer for digital cameras," *arXiv preprint arXiv:1709.01118*, 2017. 2

[43] Q. Chen, J. Xu, and V. Koltun, "Fast image processing with fully-convolutional networks," in *IEEE International Conference on Computer Vision*, 2017. 2

[44] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[45] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems*, 2016. 3

[46] W. Ren and X. Cao, "Deep video dehazing," in *Pacific Rim Conference on Multimedia*, 2017, pp. 14–24. 3

[47] W. Ren, J. Zhang, X. Xu, L. Ma, X. Cao, G. Meng, and W. Liu, "Deep video dehazing with semantic segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 4, pp. 1895–1908, 2019. 3

[48] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3

[49] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[50] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang, "Deep image harmonization," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[51] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 3

[52] J. S. Ren, L. Xu, Q. Yan, and W. Sun, "Shepard convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2015. 3

[53] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016. 4, 7

[54] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *IEEE International Conference on Computer Vision*, 2017. 4

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, 2014. 5

[56] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014. 5

[57] S. Nah, T. H. Kim, and K. M. Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 8

[58] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3253–3261. 8

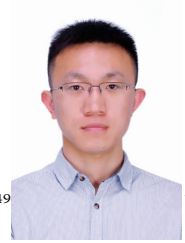
[59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing & Computer-assisted Intervention*, 2015. 9, 10

Wenqi Ren is an Assistant Professor in Institute of Information Engineering, Chinese Academy of Sciences, China. He received his Ph.D. degree from Tianjin University, Tianjin, China, in 2017. During 2015 to 2016, he was supported by China Scholarship Council and working with Prof. Ming-Husan Yang as a joint-training Ph.D. student in the Electrical Engineering and Computer Science Department, at the University of California at Merced. He received Tencent Rhino Bird Elite Graduate Program Scholarship in 2017, MSRA Star Track Program in 2018. His research interests include image processing and related high-level vision problems.

Sifei Liu is a research scientist at Nvidia Research in Santa Clara, US. She obtained her PhD at University of California Merced, department of EECS, advised by Professor Ming-Hsuan Yang. Her research interests lie in computer vision (low-level vision, semantic segmentation and 3D scene understanding), and deep learning (graph-structured building blocks). She has worked as an intern student in Baidu IDL from 2013 to 2014, multimedia lab in CUHK in 2015, and NVIDIA research in 2017. She is also the recipient of the Baidu Graduate Fellowship in 2013, NVAIL pioneering research award in 2017.

Lin Ma received the B.E. and M.E. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006 and 2008, respectively, and the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2013. He was a Researcher with the Huawei Noahs Ark Laboratory, Hong Kong, from 2013 to 2016. He is currently a Principal Researcher with the Tencent AI Laboratory, Shenzhen, China. His current research interests lie in the areas of computer vision and multimodal deep learning, specifically for image and language, image/video understanding, and quality assessment. Dr. Ma was a recipient of the Microsoft Research Asia Fellowship in 2011. He received the Best Paper Award from the Pacific-Rim Conference on Multimedia in 2008. He was the Finalist of the HKIS Young Scientist Award in engineering science in 2012.

Qianqian Xu received the B.S. degree in computer science from China University of Mining and Technology in 2007 and the Ph.D. degree in computer science from University of Chinese Academy of Sciences in 2013. She is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her research interests include statistical machine learning, with applications in multimedia and computer vision. She has authored or coauthored 20+ academic papers in prestigious international journals and conferences, among which she has published 6 full papers with the first author's identity in ACM Multimedia. She served as member of professional committee of CAAI, and member of online program committee of VALSE, etc.



Xiangyu Xu is a research scientist of SenseTime, Beijing. He received the Ph.D. degree in the Department of Electronic Engineering at Tsinghua University, China, in 2018. He was a joint-training Ph.D. student in the School of Electrical Engineering and Computer Science at University of California, Merced, CA, USA from 2015 to 2017. His research interest includes image processing, low-level vision and deep learning.

Xiaochun Cao received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, since 2012. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He has authored and coauthored more than 120 journal and conference papers. Prof. Cao is a Fellow of the IET. He is on the Editorial Board of the IEEE TRANSACTIONS OF IMAGE PROCESSING. His dissertation was nominated for the University of Central Floridas university-level Outstanding Dissertation Award. In 2004 and 2010, he was the recipient of the Piero Zamperoni Best Student Paper Award at the International Conference on Pattern Recognition.

Junping Du is a Professor at Beijing University of Posts and Telecommunications. Her research interests include social network analysis and searching, multi-source information fusion and big data mining. Prof. Du has served as an Associate Editor of International Journal of Intelligent Information Processing, is also in the Editorial Board of Computer Simulation and CAAI Transactions on Intelligent Systems. She has published more than 300 papers in international journals and conferences such as IEEE TKDE, TPAMI, and CVPR. She has served in program committee in BigComp, ICHS, CISC, etc. She is the Fellow of CCF.

Ming-hsuan Yang is a professor in Electrical Engineering and Computer Science at University of California, Merced. He received the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 2000. Yang has served as an associate editor of IEEE Transactions on Pattern Analysis and Machine Intelligence, International Journal of Computer Vision, Computer Vision and Image Understanding, Image and Vision Computing and Journal of Artificial Intelligence Research. He received the NSF CAREER award in 2012, Senate Award for Distinguished Early Career Research at UC Merced in 2011, and Google Faculty Award in 2009. He is a Fellow of the IEEE.