Lightening Network for Low-Light Image Enhancement

Article *in* IEEE Transactions on Image Processing · July 2020

DOI: 10.1109/TIP.2020.3008396

CITATIONS
232

READS
3,463

4 authors, including:

Li-Wen Wang
Tencent
37 PUBLICATIONS **1,332** CITATIONS
SEE PROFILE

Zhi-Song Liu
Lappeenranta – Lahti University of Technology LUT
65 PUBLICATIONS **1,470** CITATIONS
SEE PROFILE

Wan-Chi Siu
The Hong Kong Polytechnic University
596 PUBLICATIONS **8,568** CITATIONS
SEE PROFILE

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.2020.3008396, IEEE Transactions on Image Processing

Lightening Network for Low-light Image Enhancement

Li-Wen Wang, Zhi-Song Liu, Wan-Chi Siu, *Life-FIEEE*, and Daniel Pak-Kong Lun, *SMIEEE*
The Hong Kong Polytechnic University, Hong Kong SAR, China
{liwen.wang, zhisong.liu}@connect.polyu.hk, {enwcsiu, enpklun}@polyu.edu.hk

***Abstract*— Low-light image enhancement is a challenging task that has attracted considerable attention. Pictures taken in low-light conditions often have bad visual quality. To address the problem, we regard the low-light enhancement as a residual learning problem that is to estimate the residual between low- and normal-light images. In this paper, we propose a novel Deep Lightening Network (DLN) that benefits from the recent development of Convolutional Neural Networks (CNNs). The proposed DLN consists of several Lightening Back-Projection (LBP) blocks. The LBPs perform lightening and darkening processes iteratively to learn the residual for normal-light estimations. To effectively utilize the local and global features, we also propose a Feature Aggregation (FA) block that adaptively fuses the results of different LBPs. We evaluate the proposed method on different datasets. Numerical results show that our proposed DLN approach outperforms other methods under both objective and subjective metrics.**

***Keywords*— low-light image enhancement, image processing, deep learning.**

I. Introduction

Taking photos is one of the most popular and convenient ways to record memorable moments of our life. Images taken in low-light conditions are usually very dim. This makes us difficult to recognize the scene or object. However, often it is inevitable to take photos in low-light conditions. To obtain high-visibility images in the low-light conditions, we can adopt three solutions. 1) To use flash: It is a direct way to solve the problem. However, it is not allowed in some public areas, such as the museum, cinema, and exhibition hall. 2) To increase the ISO (sensitivity of the sensor): This method could increase the visibility of dark areas, but higher ISO will also bring more noise to the image, and the normal-light area will easily face the overexposure problem. 3) To take a photo with longer exposure time: Capturing an image with longer exposures allows more light that enlightens the dark area. Nevertheless, long-time exposure may blur the image if there is camera shake or fast-moving objects.

A large number of conventional approaches have been proposed to mitigate the degradation caused by low-light conditions. Histogram Equalization (HE) [1, 2] counts the frequency of the pixel values. By rearranging the pixels to obey uniform distribution, it improves the dynamic range (i.e., better visibility) of the low-light image. Retinex-based methods [3] regard one image as a combination of illumination and reflectance, where the reflectance is an inherent attribute of the scene that is unchangeable in different lighting conditions, and the illumination maps store the differences between the low- and normal-light images. The Retinex-based methods enhance the illumination map of the low-light image to estimate the corresponding normal-light image.

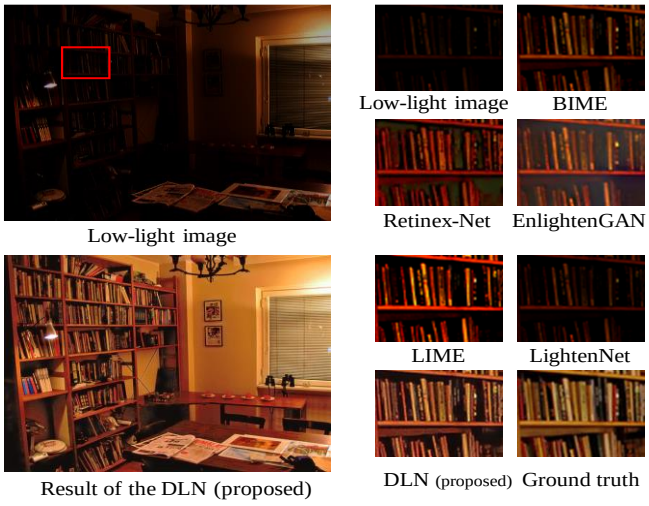

Figure 1. Effect of our proposed method

Other methods adopt dehazing theory [4, 5] that decomposes the low-light image to ambient light, refraction, and scene information. Refining the refraction map can also enhance the visibility of low-light images.

Convolutional Neural Networks (CNNs) have achieved impressive results in many tasks, such as image classification [6], semantic segmentation [7], super-resolution [8], and object detection [9]. Compared with conventional approaches, the CNNs have better feature representation that benefits from the large dataset and powerful computational ability. For CNNs, the information extracted from the shallow layers has detailed local information (like edge, texture), while deep layers have large receptive fields that can obtain more global features (like complex texture and shape) [10]. The CNNs tend to have more convolutional layers and complex structures to obtain more powerful learning abilities [11-13].

The low-light enhancement can be regarded as an image restoration task. Image Super-Resolution (SR) is one of the similar topics, which reconstructs a high-resolution (HR) image from a low-solution (LR) image of different scales. Some SR networks adopt an end-to-end structure that minimizes the mean squared error between the reconstructed SR and HR images [14-16]. Other approaches add Back-Projection structures that iteratively up- and down-sampling the LR images. It improves the efficiency of the network that is widely used in the field [8, 17, 18]. For example, Deep Back-Projection Network (DBPN) [19] approach has several BP stages that iteratively reconstruct the SR image. Back Projection and Residual Network (BPRN) [18] refines the DBPN structure by injecting the advantages of Residual Network structure. Hierarchical Back-Projection Network

(HBPN) [8] investigates the benefits of Hour-Glass and weighting structures to enhance the BPRN.

Recent literature shows that the CNN technology also benefits the low-light image enhancement. Some approaches (like Retinex-Net [20], LightenNet [21]) are based on the Retinex theory that contains two CNNs: One network decomposes the low-light image into illumination and reflectance, where reflectance is an inherent attribute of the scene which is unchangeable in different light conditions. The other network works as an enhancer to refine the illumination map of the low-light image. However, the definitions of ground-truth illumination and reflectance are not clear, which makes the decomposition difficult. Another problem is that these CNN-based approaches make use of shallow CNN structures that have few trainable parameters, which leads to a considerable limitation on the performance. For example, Retinex-Net [20] has only seven convolutional layers in the decomposition network, and LightenNet [21] has four convolutional layers only. It is obvious that the deep learning for low-light enhancement is still in its infancy stage. Some other approaches use Generative Adversarial Networks (GANs) that regard the low-light enhancement as a domain transfer learning task by finding the mapping between low- and normal-light domains (e.g. EnlightenGAN [22]). Each GAN has a generator and a discriminator, where the generator estimates normal-light images from the low-light ones, while the discriminator constrains the visual quality of the estimations and tries to distinguish the estimations from real normal-light images. However, the generator may collapse to a setting where it always outputs the same settings that are difficult for the discriminator to distinguish. In addition, the two models need to be trained simultaneously, but they have completely opposite targets that make it difficult to obtain the desired output [23].

Although considerable research has been devoted to apply the CNNs for low-light enhancement, less effort is being made to investigate new and suitable structures for the task. Recently, the use of back-projection block has shown outstanding performance in the image restoration field (e.g., image Super-Resolution(SR)). Based on the idea of enhancing the image iteratively, we proposed a novel CNN structure (i.e., the Deep Lightening Network (DLN)) that achieves remarkable enhancement for the low-light image, as shown in Fig. 1. Let us highlight the novelty of our proposed method as follows:

- **Interactive Low-light Enhancement**: We resolve the low-light enhancement through a residual learning model that estimates the residual between the low- and normal-light images. The model has an interactive factor that controls the power of the low-light enhancement. More details can be found in Section III-A.
- **Deep Lightening Network (DLN)**: We propose a novel DLN approach based on our residual model to enhance the low-light image in an end-to-end way. It contains several lightening blocks (see LBPs in Figure 2) that enhance the low-light image accumulatively. Our DLN is compared with several state-of-the-art approaches through comprehensive experiments. The results show that our proposed DLN outperforms all other methods in both subjective and objective measures.
- **Lightening Back-Projection (LBP)**: Based on the idea of enhancing the low-light image iteratively, we propose a LBP block that iteratively lightens and darkens the low-light image to learn the residual for low-light enhancement. It is the first work that successfully introduces a new back-projection structure for low-light enhancement. More details can be found in Section III-C.
- **Feature Aggregation (FA):** Both global and local features are useful for low-light enhancement. We propose a FA block that aggregates the results from different lightening stages and provides more informative features for the following lightening process. More details can be found in Section III-D.

The rest of the paper is organized as follows: Section II presents a brief review of some related works. Section III models the low-light enhancement firstly, and then introduces our proposed DLN method. Section IV shows experimental results, and Section V concludes the paper.

II. Related Works

Back Projection (BP): BP technique has been widely used in image Super-Resolution (SR) tasks, that initially utilizes multiple low-solution (LR) images to predict one SR image. The Deep Back-Projection Network (DBPN) [17] comes up with refining the quality of SR by using BP blocks iteratively, which minimizes the loss between LR and down-sampled SR images. The BP block can be described as follows: Assume that we have obtained an immediate SR image (denoted as $\hat{\mathbf{Y}}_t \in \mathbb{R}^{H\times W\times 3}$, where H, W and 3 mean the height, width and RGB channels, respectively. The symbol t denotes the iteration index). Firstly, BP down-samples the SR image and calculates the residual with the LR image (denoted as $\mathbf{X} \in \mathbb{R}^{H'\times W'\times 3}$, where H' and W' mean the down-sampled height and width, respectively). Then, it up-samples the residual to obtain the residual between SR image $\hat{\mathbf{Y}}_t$ and the ground-truth high-resolution (HR) image. By adding the residual with a balance coefficient $\lambda \in \mathbb{R}$, the SR image $\hat{\mathbf{Y}}_t$ can be refined as $\hat{\mathbf{Y}}_{t+1}$. Mathematically, the BP block is described as:

$$\hat{\mathbf{Y}}_{t+1} = \hat{\mathbf{Y}}_t + \lambda U(\mathbf{X} - D(\hat{\mathbf{Y}}_t)) \tag{1}$$

where $D(\cdot)$ and $U(\cdot)$ represent the down-sampling and up-sampling operations separately.

The approach can be improved by adding two weighting coefficients ($\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$, as shown in Eqn. 2) to form an enhanced BP block [8], which makes use of the residual information more efficiently.

$$\hat{\mathbf{Y}}_{t+1} = \beta\hat{\mathbf{Y}}_t + \lambda U(\alpha\mathbf{X} - D(\hat{\mathbf{Y}}_t)) \tag{2}$$

For the task of image SR, the LR images may lose some detail information after the down-sampling process. The SR algorithms up sample the LR images by estimating the lost details. The low-light enhancement is different from the image SR task, which takes no account of scale changes but refining the illumination conditions of the low-light images. Therefore, using this BP block could not resolve the low-light problem. In this paper, based on the idea of the BP process, we propose a novel Lightening Back-projection (LBP) block that focuses on increasing the dynamic range of the low-light image. More details can be found in Section III-C.

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.2020.3008396, IEEE Transactions on Image Processing

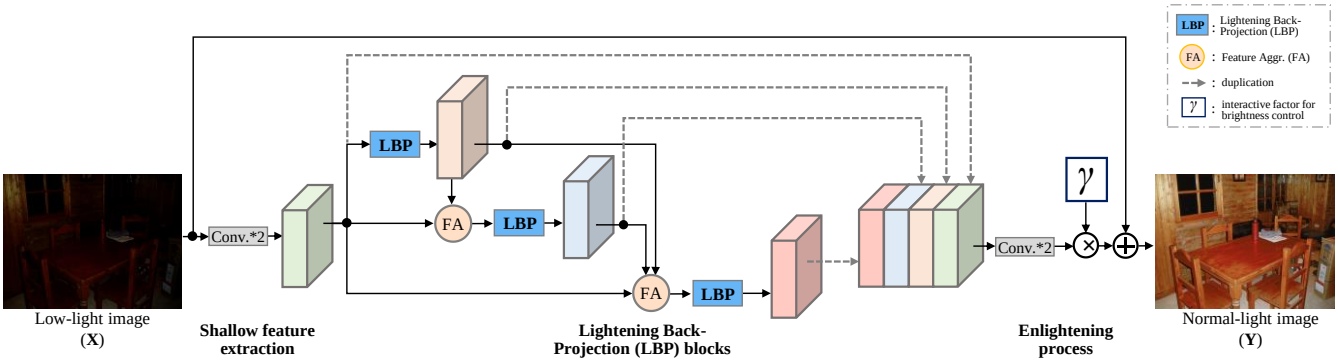

Figure 2. Architecture of Deep Lighten Network (DLN). The rectangles and cubes denote the operations and feature maps respectively. LBP represents the Lighten Back-Project (LBP) block (see Section III-C for details). FA denotes the feature aggregation block (see Section III-D for details). γ is interactive factor of brightness control (see Section III-A for details).

Feature Recalibration: CNN consists of several convolutional layers, and each layer has a set of trainable filters that express different local spatial patterns. Then, CNN can produce image representations by capturing hierarchical patterns from the filters. Some approaches strengthen the representation power by investigating the spatial encodings from different sizes of the filters, like the Inception family [12, 24, 25]. Other approaches, like Squeeze-and-Excitation (SE) block [26], improve the representation ability by seeking the relationship on the channels. To investigate the interdependencies between channels, it firstly uses average global pooling at each channel to squeeze the global spatial information into channel-wise descriptors (denoted as $F_{sq}(\cdot)$ in Eqn. 3).

$$F_{sq}(\mathbf{U}_c) = \frac{1}{\mathrm{H}\times\mathrm{W}}\sum_{i=1}^{\mathrm{H}}\sum_{j=1}^{\mathrm{W}} u_c(i,j) \tag{3}$$

where $\mathbf{U}_c \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}}$ denotes the feature map with channel c. The symbol $c \in \mathbb{R}$ denotes the c-th channel of the feature map, and $u_c(i,j) \in \mathbb{R}$ represents the attribute value at position (i,j) of the feature map.

Then, the channel-wise descriptors are sent to a shallow neural network to model the interdependencies among the channels. It can be regarded as a feature selection process that assigns different weights (denoted as $w \in \mathbb{R}^{\mathrm{C}}$) to different feature channels. Finally, it recalibrates the feature maps by multiplying the weights with the corresponding maps (denoted as $F_{scale}(\cdot)$ as shown in Eqn. 4).

$$F_{scale}(\mathbf{U}_c, w_c) = w_c \cdot \mathbf{U}_c \tag{4}$$

where $w_c \in \mathbb{R}$ is the estimated weight for the channel $\mathbf{U}_c$.

Making use of the hierarchical structure, CNNs have inherent multi-scale feature representations, where the features extracted from the shallow layers usually contain detailed information (like edge and texture), and the features extracted from deep layers provide global components (like complex texture and shape). For low-light enhancement, both global and local information is essential. The global information is helpful for the evaluation of the illumination condition, and the local features benefit the detail restoration. Therefore, fusing both local and global information can construct wealth features for the following process.

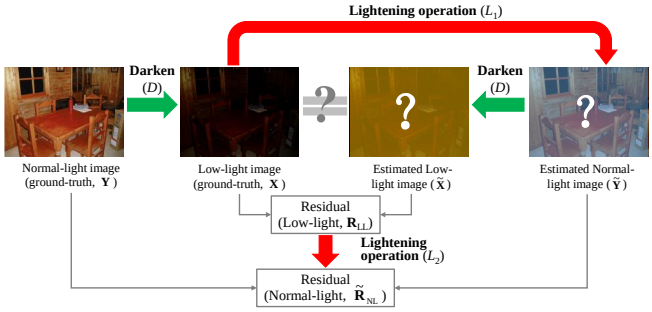

Figure 3. Relation between the low- and normal-light images.

Nevertheless, features from different layers play distinct roles in the feature representation. Stacking the feature maps may simply lose some representation power. Hence, further investigation for the channel-wise dependencies is needed. However, very few papers in the literature focus on seeking for a better representation from different layers. Based on the idea of squeeze-and-excitation, we propose a Feature Aggregation (FA) block that strengthens the feature representation power from multiple intermediate layers, which fuses both spatial and channel-wise information to the same block. More details can be found in Section III-D.

III. The Proposed Deep Lightening Network

Firstly, we model the low-light enhancement as a residual learning task. Then, we present our proposed Deep Lightening Network (DLN) that learns the residual for the low-light enhancement.

A. Assumption: Residual Learning

Single low-light image enhancement is a fundamental low-level vision problem where the aim is to reconstruct a normal-light (NL) image $\mathbf{Y} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times3}$ from a low-light (LL) image $\mathbf{X} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times3}$. However, it is difficult to get paired LL-NL images as there may not be a unique or well-defined ground-truth NL image given a LL image. Therefore, instead of learning the mapping function between the LL and NL images directly, we model the problem as a residual learning task, and the assumption is shown below:

$$\mathbf{Y} = \mathbf{X} + \gamma P(\mathbf{X}) - \mathbf{n} \tag{5}$$

where $P(\cdot)$ denotes the enhancing operator that estimates the residual between the NL and LL images. We introduce an interactive factor $\gamma \in \mathbb{R}$ that controls the lightening power of the low-light enhancement (its effect is shown in Fig. 7). The symbol $\mathbf{n} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times3}$ represents the noise to be removed. To simplify the low-light enhancement task, the noise term is ignored in this paper. Then, the low-light enhancement is to find an enhancing operator $P(\cdot)$, which can be learned by a CNN structure. The optimization of the CNN is formulated as:

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.2020.3008396, IEEE Transactions on Image Processing

$$P = \text{argmin}_E(\|\mathbf{Y} - (\mathbf{X} + \gamma \cdot P(\mathbf{X}))\|_2 + \lambda \cdot \Omega(P)) \quad (6)$$

where $\|.\|_2$ represents the L2 norm, and $\lambda \in \mathbb{R}$ is a factor to balance the regularization term $\Omega(P)$.

B. Deep Lightening Network (DLN)

Fig. 2 illustrates the architecture of our proposed Deep Lightening Network (DLN). It consists of three parts: shallow feature extraction, Lightening Back-Projection (LBP) blocks and enlightening process. The DLN takes the LL image as the input. It firstly enters into the shallow feature extraction part that consists of two convolutional layers (the Conv.*2 at the left side of Fig. 2), where each layer has 64 3-by-3 filters with stride of 1, padding of 1. Then, the multiple LBPs (with feature aggregation (FA) blocks) scheme starts to enhance the LL image accumulatively. Next, the enlightening process receives the results from LBPs and estimates the residual to the NL image by two convolutional layers (the Conv.*2 at the right side of Fig. 2. The filter size is 3*3 with stride of 1, padding of 1). Finally, the LL image is enhanced by adding the residual with the interactive factor γ. We will show the details of LBP and FA blocks in the remaining parts of this section.

C. Lighten Back-Projection (LBP)

Based on the back-projection theory (as shown in Fig. 3), a low-light (LL) image ($\mathbf{X}$) can be obtained from its normal-light (NL) version ($\mathbf{Y}$) through a darkening operation (D, see the left green arrow in Fig. 3). The objective of LL enhancement is to find a lightening operation (L_1), which predicts the NL image ($\tilde{\mathbf{Y}} \in \mathbb{R}^{H\times W\times 3}$) from the observed LL image ($\mathbf{X}$) (see the top red arrow in Fig. 3). Objectively, we can also estimate a version of the LL image ($\tilde{\mathbf{X}} \in \mathbb{R}^{H\times W\times 3}$) from the estimated NL one ($\tilde{\mathbf{Y}}$) through the darkening operation (D, see the right green arrow in Fig. 3). If the lightening (L_1) and darkening operations (D) are in an ideal situation, the ground-truth ($\mathbf{X}$) and estimated ($\tilde{\mathbf{X}}$) LL images will be the same. In real condition, their difference (a residual term $\mathbf{R}_{\text{LL}} \in \mathbb{R}^{H\times W\times 3}$, for $\mathbf{R}_{\text{LL}} = \mathbf{X} - \tilde{\mathbf{X}}$) indicates the weakness of the lightening (L_1) and darkening (D) operations. Based on the residual information ($\mathbf{R}_{\text{LL}}$), it can estimate the residual ($\tilde{\mathbf{R}}_{\text{NL}} \in \mathbb{R}^{H\times W\times 3}$, for $\tilde{\mathbf{R}}_{\text{NL}} \approx \mathbf{Y} - \tilde{\mathbf{Y}}$) in the NL domain through a lightening operation (L_2). Finally, the intermediate NL estimation ($\tilde{\mathbf{Y}}$) can be refined by adding the residual $\tilde{\mathbf{R}}_{\text{NL}}$ to $\tilde{\mathbf{Y}}$, i.e., $\hat{\mathbf{Y}} = \tilde{\mathbf{Y}} + \tilde{\mathbf{R}}_{\text{NL}}$, where the term $\hat{\mathbf{Y}} \in \mathbb{R}^{H\times W\times 3}$ is the refined NL estimation.

Accordingly, we propose a Lighten Back-Projection (LBP) block that is shown in Fig. 4, where each LBP block consists of two lightening, and one darkening operator. The LL image ($\mathbf{X}$) makes use of a Lightening operator L_1 to estimate a NL image ($\tilde{\mathbf{Y}}$). Next, a Darkening operator (D) predicts the LL image ($\tilde{\mathbf{X}}$) from the estimated $\tilde{\mathbf{Y}}$. For the LL image, the estimated ($\tilde{\mathbf{X}}$) should be close to its ground truth ($\mathbf{X}$). Then, it calculates the difference between $\tilde{\mathbf{X}}$ and $\mathbf{X}$, i.e., the residual ($\mathbf{R}_{\text{LL}}$). Similarly, for the residual ($\mathbf{R}_{\text{LL}}$), another Lightening operator (L_2) is used to estimate the residual ($\tilde{\mathbf{R}}_{\text{NL}}$) under NL conditions. The final estimation for the NL image ($\hat{\mathbf{Y}}$) is obtained by adding the NL estimation ($\tilde{\mathbf{Y}}$) and its residual $\tilde{\mathbf{R}}_{\text{NL}}$.

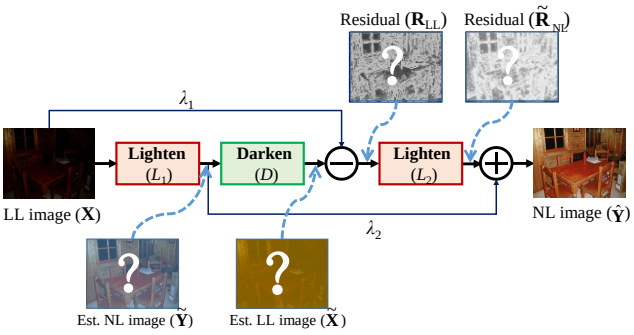

Figure 4. Structure of Lighten Back-Projection (LBP). Details of the Lighten and Darken operations are shown in Fig. 5

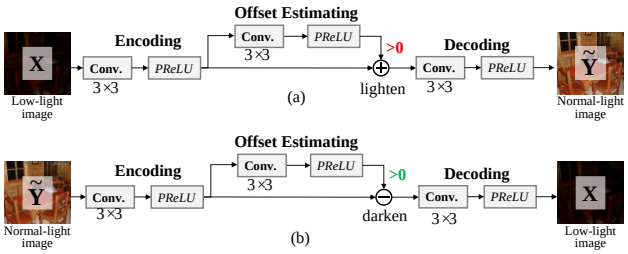

Figure 5. Structure of the lightening(a) and darkening(b) operations

As we mentioned above, different from other approaches that directly learn the mapping function between LL and NL images, the proposed LBP blocks iteratively lightening and darkening the LL image to learn the residual term ($\tilde{\mathbf{R}}_{\text{NL}}$) for a better reconstruction. The whole procedure of LBP can be formulated as:

$$\hat{\mathbf{Y}} = \lambda_2 L_1(\mathbf{X}) + L_2(D(L_1(\mathbf{X})) - \lambda_1 \mathbf{X}) \quad (7)$$

where $\lambda_1 \in \mathbb{R}$ and $\lambda_2 \in \mathbb{R}$ are two weights to balance the residual updating.

The key parts of the LBP block are the Lightening and Darkening operations. For the LL and NL images, the differences lie in the pixel magnitude of the image, i.e., the LL image usually has lower pixel values and narrower dynamic ranges compared with the NL image. Therefore, increasing or decreasing the pixel values with appropriate offsets can realize the lightening or darkening operations.

Fig. 5 shows our proposed lightening and darkening operations, where each operator consists of three parts: encoding, offset estimating, and decoding process. To take the lightening operation for example (see Fig. 5 (a)), the LL image (actually, it is the features of the LL image) firstly enters into an "Encoding" structure to extract representative features from the low-light image by using a convolutional block (Conv.+PReLU, which reduces the number of feature channels from 64 to 32). As we mentioned before, the lightening operation is to increase the mean values of the image. The "Offset" structure adopts a convolutional layer to learn the differences between the LL and NL images. Consider that the NL images usually have larger pixel values compared with the LL images. Note that the PReLU activation layer has the effect to remove the negative values of the offset. Then, adding the offset to the LL image can increase the pixel values of the LL image, i.e., lightening the LL image. Subsequently, the "Decoding" process is conducted to reconstruct the NL image (actually, increase the number of the feature channels from 32 to 64). Similarly, the darkening operator estimates the offset and performs the subtraction to darkening the images (see Fig. 5 (b)).

D. Feature Aggregation (FA)

As shown in Fig. 2, the DLN has several short connections among the LBPs, which allows to propagate

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.2020.3008396, IEEE Transactions on Image Processing

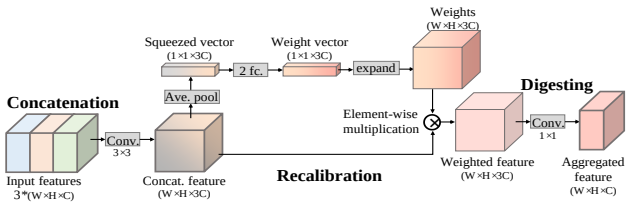

Figure 6. Structure of three-input Feature Aggregation (FA) block

features from the former to the latter LBPs. To use the features more effectively, we propose a feature aggregation (FA) block that strengthens the feature representation power based on multiple intermediate results. The first FA block in Fig. 2 on the left has two inputs which fuses information from two feature maps, while the second FA block in Fig. 2 on the right fuses three input feature maps. Let us use the second FA block as an example (as shown in Fig. 6) which receives three feature maps, and each map has a size of W×H×C, where W, H, and C denote the width, height, and the number of channels of the feature map separately. The FA block consists of three parts: feature concatenation, recalibration, and digesting process, as described below.

Feature concatenation: For a CNN network, the shallow layers extract the feature maps that contain detailed information. After the process of several layers, the neurons have larger receptive fields that extract more global information. Therefore, the filters of different layers can investigate the information on different sizes of spatial regions. As we mentioned above, for the low-light enhancement task, both local and global information are essential, because we need global information to evaluate the light condition of the whole image and the local features to refine the details. The FA block takes multiple feature maps (three cubes with different colors in Fig. 6) that contain different spatial information as the input. It concatenates them together (with size: W×H×3C) and captures the spatial correlations in different scales through a convolutional layer, where the filter size is 3, with stride of 1 and padding of 1.

Recalibration: Each channel of the feature map stores the information of a type of spatial pattern that is extracted by the convolutional filter. Based on the idea of constructing informative features by fusing the channel-wise information [26], we recalibrate the concatenated feature map by giving weights to different channels. As shown in Fig. 6, the recalibration process contains a weighting branch and a short connection. For each channel (size: W×H), the weighting branch squeezes the information into a single value through global average pooling ("Ave. pool" in Fig. 6, mathematically the process can be written as F_{sq} in the Eqn. 3). Then, the feature map (size: W×H×3C) can be described by a squeezed vector with the size of 1×1×3C, and each value represents the information for one channel. To investigate the channel-wised dependency, we make use of two fully-connected (fc) layers ("2 fc." in Fig. 6, where the first layer consists of C/16 neurons and the second layer has C neurons) to assign weights for different feature channels, i.e., it estimates a weight vector (size: 1×1×3C), each attribute of which stores the weight for each channel. Next, it expands the weights at the width-height plane, and this changes the dimension to W×H×3C. Finally, the representational ability of the feature map is improved by multiplying the weights to the corresponding features (see "Element-wise multiplication" in Fig. 6, and mathematically the process can be written as F_{scale} in Eqn. 4). Note again that the recalibration process investigates channel-wise dependencies of the concatenated feature maps.

Digesting: Usually, the recalibration process has the effects to make the key features have large weights, that are more important for the following process. The digesting block further improves the representation ability of the weighted features through a one-by-one convolutional layer ("Conv." on the right in Fig. 6), which reduces the channels from W×H×3C to W×H×C.

E. Loss Function

Given a LL image, we can estimate its corresponding NL image through the network. We consider the low-light enhancement as a supervised learning task where for each LL image, there is a corresponding NL image as the training target. Therefore, we define the loss function as shown in Eqn. 8. It consists of two parts: $Loss_{struct}$ which measures the structure similarity, and $Loss_{TV}$ which constrains the smoothness that works as a regularization term.

$$Loss_{dif}(\hat{\mathbf{Y}}, \mathbf{Y}) = Loss_{struct}(\hat{\mathbf{Y}}, \mathbf{Y}) + \lambda\, Loss_{TV}(\hat{\mathbf{Y}}) \tag{8}$$

where $\lambda \in \mathbb{R}$ is the balance coefficient (note: we used 0.001 in our experiments).

Structure Similarity: Images captured in the low-light condition usually have structure distortion problems (like blur effect) that are visually salient [27]. MAE and MSE losses average all pixel-wise differences that cannot competently handle the problem. In order to improve the quality of the estimation both qualitatively and quantitatively, we use the Structure Similarity (SSIM, definition is shown in Eqn. 9) [28] as the evaluation metric that gives further consideration for structure similarity. The value of SSIM ranges from 0 to 1, and larger value means better similarity. Therefore, we define the structure loss as $Loss_{struct}(\hat{\mathbf{Y}}, \mathbf{Y}) = 1 - SSIM(\hat{\mathbf{Y}}, \mathbf{Y})$.

$$SSIM(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \cdot \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{9}$$

where $x \in \mathbb{R}^{H\times W\times 3}$, $y \in \mathbb{R}^{H\times W\times 3}$ denote the images to be measured. The terms $\mu_x \in \mathbb{R}$ and $\mu_y \in \mathbb{R}$ are the mean values of the two images. The symbols $\delta_x \in \mathbb{R}$ and $\delta_y \in \mathbb{R}$ represent the variances of the images. The terms $c_1 \in \mathbb{R}$ and $c_2 \in \mathbb{R}$ are two constants to prevent the denominator be zero (c_1=0.0001, c_2=0.0009 were used in our experimental work, which are the same settings as those in [29]).

The Constraint of Smoothness: the NL estimation may have inconstant illumination and noises that decrease the visual quality. Total Variation (TV) is used in our work as smoothness prior that minimizes the gradient of the whole image. The definition of TV loss is shown below:

$$Loss_{TV}(\mathbf{P}) = \sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{3}\left(\frac{(p_{i,j,k} - p_{i+1,j,k})^2}{3H(W-1)} + \frac{(p_{i,j,k} - p_{i,j+1,k})^2}{3(H-1)W}\right) \tag{10}$$

where $\mathbf{P} \in \mathbb{R}^{H\times W\times 3}$ denotes the image to be measured. The symbol $p \in \mathbb{R}$ represents the pixel values. The terms i and j are the indexes of the pixels.

IV. Experiments

There is no objective evaluation method for the light-condition measurement that makes it difficult to evaluate the performance of different low-light enhancement methods. We believe that the enhanced LL image should be close to the ground-truth NL image. Therefore, we adopt Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity

(SSIM) [28], which are widely used in image restoration field to measure the quality of the estimation. Subjectively, we will present visualization results for comparison. We will also compare our proposed DLN method with existing approaches that were implemented through their publicly available codes, including: conventional methods (HE [1], BIMEF [30], LIME [31]), CNN-based methods (LightenNet [21], LLNet [32], Retinex-Net [20]), and GAN-based methods (EnlightenGAN [22]) on a synthetic dataset (images were generated by mathematical approach, as described in Section IV-A). To further evaluate the generalization ability of the proposed work, we have also tested it on a real dataset (images were captured in the real situation). Program codes will be released in the same month of the publication date.

A. Implementation Details

Low-light Image Synthesis: CNN has a large number of trainable parameters that need a huge training dataset (i.e., the LL-NL image pairs) for the training process. However, it is difficult to capture the LL-NL images of the same scene at the same time. The NL images usually contain more information and less noise as compared with the LL ones. It is feasible to synthesize the LL images from the NL images. Following the analysis in [33], a LL image can be simulated from a NL image through the following simulation equation:

$$\overline{\mathrm{X}}_{i,j}^{(c)} = \beta \left(\alpha \, \mathrm{Y}_{i,j}^{(c)} \right)^{\gamma} \tag{11}$$

where $\overline{\mathrm{X}}_{i,j}^{(c)} \in \mathbb{R}$ represents the pixel of the simulated LL image. The terms i and j are the locations of the pixels. The term $c \in \{\mathrm{R}, \mathrm{G}, \mathrm{B}\}$ denotes the R, G, or B channel of the image. The pixel $\mathrm{Y}_{i,j}^{(c)} \in \mathbb{R}$ is from the NL image that is compressed to [0, 1]. The symbols α, β and $\gamma \in \mathbb{R}$ follow the uniform distribution, i.e., $\alpha \sim U(0.9, 1)$, $\beta \sim U(0.5, 1)$ and $\gamma \sim U(1.5, 5)$, which control the effect of low-light simulation (the settings are the same as [33]). γ is a non-linear element of the simulation equation that has the effect of local illumination transformation. It gives a stronger dark effect to the low-light regions of the NL images, which fits very well the real situation (low-light regions of the NL images are usually darker than other regions in the corresponding LL images).

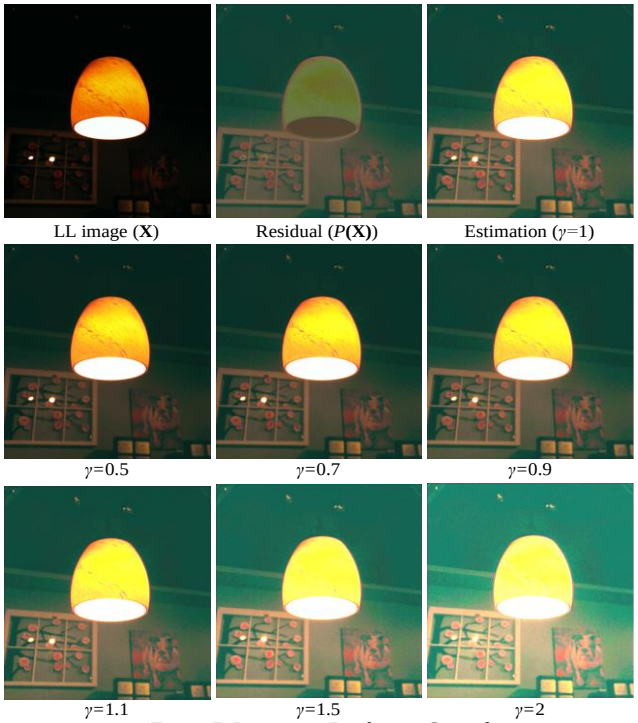

Figure 7. Interactive Brightness Control

Training Settings: Images in PASCAL VOC 2007 dataset have good visual quality that were initially used for image object detection. We used all 9,963 images in PASCAL VOC 2007 dataset (train + validation + test) as the ground-truth NL images, and assumed that all of them have ideal illumination conditions (then, we can set γ=1 in Eqn. 5 at the training stage). The images were resized through the bicubic method, making the shorter side of the images have 384 pixels, where the original aspect ratios of the images are kept. Then, we simulated the LL images from these resized images based on Eqn. 11 with data augmentations (like reducing the contrast, color, etc., randomly) by using the Pillow [34] package, where α, β, and γ are randomly selected from their ranges for each image.

Let us refer to the architecture of our DLN in Fig. 2 and Section III. We randomly initialized the weights of DLN as in [35]. The Adam method was adopted to optimize the parameters with the momentum equal to 0.9 and the weight decay equal to 0.0001. The learning rate was set to 0.0001 for all layers. To produce more LL-NL pairs, we randomly cropped 128*128 patch pairs from LL and NL images in the training phase. Because the network is a fully convolutional structure, a filter is shared by the whole image. In the testing phase, the testing images were processed with their original sizes. For each iteration, we set the mini-batch size as 32, and the model was trained for 100 epochs. All experiments were conducted using PyTorch on a two-GPU (NVIDIA GTX2080Ti) PC.

B. Analysis of Network Structure

We used different sets of data for the training and testing processes. Our testing dataset was obtained from 100 NL images (with abundant illumination and structure) that were selected from VOC 2012 testing dataset. Then, we used the same simulation method (as presented in Section IV-A) to generate the corresponding LL images. These LL-NL image pairs then form our testing dataset and are used for the following analysis. It is interesting to know the effect of our proposed DLN network, LBP, and FA blocks separately. We are going to investigate their effects and find the optimal settings by presenting a set of experimental results.

1) Effect of the Residual Learning

As we introduced in Section III-A, we regard the low-light enhancement as a residual-learning task where our DLN model learns the residual (**Y**-**X**) for the lightening process. In order to evaluate the effectiveness of the residual model, we have made a comparison with a direct-learning model that learns the mapping from **X** to **Y** directly (i.e. removing the "short connection (**X**)" in Fig. 2). Table I shows the results of the direct- and residual-learning models. We can find that the proposed residual-learning model achieves better PSNR and SSIM scores, which confirms that the residual-learning model is more suitable to the low-light enhancement compared

TABLE I. COMPARISON BETWEEN DIRECT- AND RESIDUAL-LEARNING MODELS

Model	PSNR	SSIM
Direct Learning	21.602	0.873
Residual Learning	**23.829**	**0.912**

with the direct-learning model. The reason is that, for the residual-learning model, the estimation preserves all the information of the LL image, and the learning process is simply to find a lightening residual. While for the direct-learning model, it needs to reconstruct the NL estimation comprehensively, which is more complicated compared with the residual-learning model. Therefore, optimizing the residual-learning model is much easier than to optimize the original, unreferenced direct-learning model.

Interactive Brightness Control: Based on Eqn. 5, the NL image can be obtained by adding a lightening residual $P(\mathbf{X})$. γ in the equation controls the power of lightening process. We can control the value of γ before the addition operator (as shown in Fig.2). Fig. 7 illustrates an example on the effect that γ was interactively adjusted with different values. For a LL image ($\mathbf{X}$, as shown in Fig. 7), the DLN can estimate the lightening residual ($P(\mathbf{X})$, as shown in Fig. 7) in the ideal lighting condition (i.e., γ=1). It can be seen from the figure that a larger γ leads to more remarkable enhancement. When γ=0.5, the LL image is lightened slightly, which makes the mural become visible. When "γ=1", the LL image is enhanced appropriately. When γ=2, the enhancement is too strong that the ceiling lamp becomes overexposed, as shown in Fig. 7. The interactive brightness control gives us the chance to control the effectiveness of the enhancement.

2) Effect of the DLN Structure

To evaluate the effect of our LBP block, let us make a comparison with the PlainNet and ResModel, where the PlainNet stacks convolutional blocks (conv.+PReLU) one by one, which is a standard CNN structure. ResNet [11] is a popular CNN structure that consists of several residual blocks. Each block has a skip connection that delivers the information from shallow layers to deep layers, which improves the performance of the deep CNNs. We stacked a set of residual blocks (the feature maps keep the same size as the input) to form a residual network, which is named ResModel in this paper. To make the comparison as fair as possible, we assigned similar parameters to the Plain net, ResModel, and DLN. It can be seen from Table II that the ResModel exceeds plain net with 0.774dB (=20.818-20.044) on PSNR and 0.072 (=0.862-0.790) on SSIM. The DLN further outperforms the ResModel with 3.011dB (=23.829-20.818) on PSNR, and 0.05 (=0.912-0.862) on SSIM. Our proposed DLN structure achieves the best PSNR and SSIM scores in the evaluation. The reason is that the structure of ResModel has the skip connections that can transmit information from the input to the output. It simplifies the learning task to the residual information which is much easier than directly learning the input-to-output mapping, such as the PlainNet. Our proposed DLN structure has more theoretical support for the low-light enhancement. The DLN has several LBP blocks that iteratively lighten and darken the image and can increase the efficiency of the residual-learning process. The experimental result shows that our DLN structure makes better use of the learning ability of CNN, and benefits the low-light enhancement.

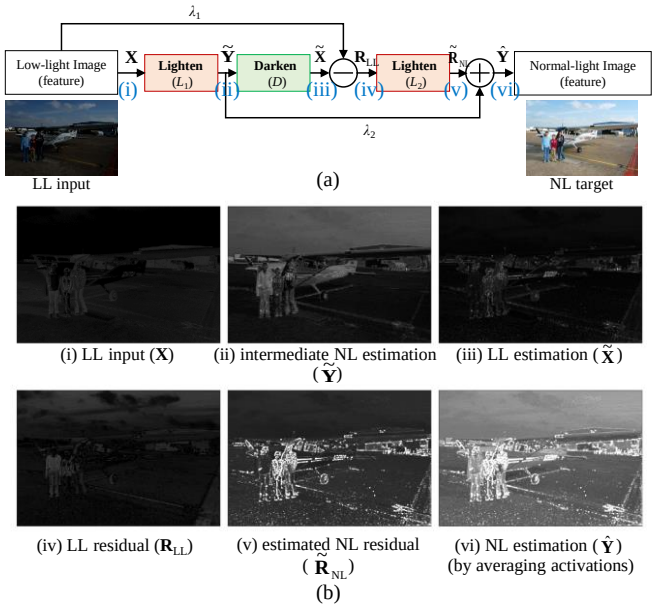

Figure 8. Intermediate results of the LBP (the top one of the DLN). The LL input and NL target of the DLN are shown at the top right corner. Visualization is made by averaging the activations of all feature channels.

TABLE II. COMPARISON OF DIFFERENT CNN STRUCTURES

Structure	PSNR	SSIM
PlainNet	20.044	0.790
ResModel	20.818	0.862
DLN	**23.829**	**0.912**

3) Effect of the Lightening and Darkening Operations

The lightening and darkening operations play important roles in the LBP blocks. The structure of this proposed lightening and darkening operation has been discussed in Section III-C, and let us call it "LBP-DLN" in the experiments. Each operation consists of three parts: encoding, offset estimation and decoding process. For the sake of comparison, we formed a modified structure that stacked three convolutional blocks one by one to form a plain structure, and let us call it "LBP-Plain" in the experiment. Table III shows the results. We can see that the proposed lightening and darkening operations (LBP-DLN) achieve better PSNR and SSIM scores, which shows the effectiveness of the proposed lightening and darkening operations.

TABLE III. COMPARISON OF DIFFERENT CNN STRUCTURES

Structure	PSNR	SSIM
LBP-Plain	23.157	0.903
LBP-DLN (proposed)	**23.829**	**0.912**

The lightening operation increases the brightness by raising the mean value of the image, while the darkening operation decreases the brightness that reduces the mean value. Let us use the "LL input" of Fig. 8(a) as an example to explain the effect of the two operations. For a trained model, we took a LBP (see Fig. 2, the top-left LBP of the DLN architecture) as an example to visualize the intermediate results, where each intermediate result (feature map) has 64 channels. We averaged all 64 channels to obtain a single map for visualization. The visualization results are shown in Fig. 8(b). Fig. 8(b)(i) gives the visualization of the LL input (the mean value is 0.0134). After the lightening operation (L_1), the map is brightened (see the sky of (ii), the mean value is 0.0553). Then, the darkening operation (D) maps the NL image back to the LL domain (as shown in (iii). Note that the sky returns to black, and the mean value is decreased to 0.0169). Subsequently, the subtraction operation finds the residual in the LL domain (as shown in (iv)), and 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.2020.3008396, IEEE Transactions on Image Processing

TABLE IV. COMPARISON ON DIFFERENT NUMBER OF LBPS

# LBP	PSNR	SSIM	Parameters	Time	Model Size
1	21.925	0.833	305k	3.84 ms	1.2MB
2	22.600	0.881	499k	5.27 ms	2.0MB
3	**23.829**	**0.912**	**700k**	**6.53 ms**	**2.8MB**
4	23.079	0.876	909k	8.35 ms	3.7MB
5	23.268	0.900	1,126k	12.37 ms	4.5MB

lightening operation (L_2) enhances it into NL residual (as shown in (v)). Finally, by adding the NL residual ($\tilde{\mathbf{R}}_{\mathrm{NL}}$) to the intermediate NL estimation ($\tilde{\mathbf{Y}}$), it can obtain the refined NL result ($\hat{\mathbf{Y}}$, see (vi) in the figure). The results confirm that the lightening or darkening operation can shift the mean values, and can increase or decrease the brightness of the input image, which exactly fits with our design.

4) Different Number of LBP Blocks

Deeper CNN has more trainable parameters that usually lead to better learning ability. We investigated the performance of the DLN with different numbers of LBPs. It can be seen from Table IV that the DLN with more LBPs has more parameters such that it needs more inference time. When the number of LBPs is less than three, using more LBPs leads to better PSNR and SSIM, which means a better lightening performance. However, when the number of LBP is larger than four, the PSNR and SSIM start to fluctuate. The reason is that the number of LBPs saturates at three, and the more LBPs makes it fluctuate between overfitting and normal tolerance situations, whilst the requirement for a larger computational power of more LBP's is also undesirable. Too many parameters may also be the cause of the overfitting problem that limits the generalization ability. Therefore, we set the number of LBPs to three to obtain the best performance.

5) Effectiveness of Feature Aggregation (FA) Block

To investigate the effect of FA block, we made a comparison on two DLNs: one removes all the FA blocks (i.e., "No FA" in Table V), and the other is the DLN that contains the FA blocks (i.e., "With FA" in Table V). We can see from Table V that using the FA block can increase the PSNR and SSIM with an extensive range. It is easy to understand that the FA blocks strengthen the feature representation by investigating the spatial and channel-wise dependencies, which are helpful to the low-light enhancement process.

TABLE V. EFFECTIVENETSS OF THE FA BLOCK

Model	PSNR	SSIM
No FA	21.887	0.875
With FA	**23.829**	**0.912**

6) Loss Function

We evaluated different loss functions for the low-light enhancement: L1-norm, L2-norm, and SSIM losses. The results are shown in Table VI. Using L1- or L2-norms achieves similar SSIM scores. Nevertheless, for the PSNR values, the estimation from L1-norm exceeds those from L2-norm with a large range (0.523dB=23.473-22.950dB). The reason is that L1- and L2-norm loss functions weigh error differently. The L2-norm exaggerates larger errors but gives small effect to small errors, while L1-norm treats all errors

TABLE VI. EFFECTIVENETSS OF THE FA BLOCK

Loss Function	PSNR	SSIM
L1-norm	23.473	0.888
L2-norm	22.950	0.885
SSIM Loss	**23.829**	**0.912**

equally. Also, the derivative of L2-norm approaches to zero when the error is tiny that hampers the training of the network (the derivative of L1-norm is always one) [36]. Therefore, the network trained with L1-norm produces better estimation with smaller errors compared with that of the L2-norm. For the SSIM loss, the trained network achieves much higher SSIM scores compared with L1-norm. It is obvious that training with SSIM loss can obtain higher SSIM scores at the testing stage, as they both desire more structure similarity. Also, the model trained by SSIM loss causes fewer errors that lead to a higher PSNR score. It confirms that using SSIM loss can benefit the training of the DLN.

7) Evaluation on the Synthetic Dataset

Table VIII shows the comparison of several existing low-light enhancement approaches on the dataset. As we mentioned before, PSNR and SSIM can work as the indices of the low-light enhancement. A larger value means the estimation has better similarity with the ground-truth reference. It can be seen from the table that the proposed DLN approach outperforms all other methods with a large extent, where it exceeds the second-best approach (LLNet) by 4.4684dB =23.829-19.145dB on PSNR and 0.12=0.912-0.792 on SSIM. It means that the estimations from DLN have better similarity to the ground-truth image, which suggests a better effort of the low-light enhancement.

TABLE VII. COMPARISON ON SYNTHETIC DATASET (**RED**: BEST; **BLUE**: THE 2ND BEST, **GREEN**: THE 3RD BEST)

Method	PSNR	SSIM
HE	15.890	0.662
BIMEF	14.943	**0.711**
LIME	15.580	0.629
LightenNet	14.317	0.600
LLNet	**19.145**	**0.792**
Retinex-Net	14.875	0.661
EnlightenGAN	**16.609**	0.682
DLN (proposed)	**23.829**	**0.912**

Fig. 9 shows a visual comparison of these methods. We can see from the figure that HE significantly improves the brightness of the LL images. However, the dynamic range is narrow which loses some information (see for example, the motorcycle in Fig. 9(a)). BIME, LIME, LightenNet, Retinex-Net, and EnlightenGAN notably improve the visual quality of the LL images. It seems that the results of LLNet and DLN have the best performance among all the methods. To further investigate the differences between the two methods, let us check the detailed reconstruction of the shoe (red rectangle in Fig. 9(a)) and the television screen (red rectangle in Fig. 9(b)) areas. We can find from Fig. 9(c) that LLNet produces blur results for the shoe, which is difficult to distinguish the brown texture, while the result of DLN shows better reconstruction. For the television area (see Fig. 9(d)), we

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.2020.3008396, IEEE Transactions on Image Processing

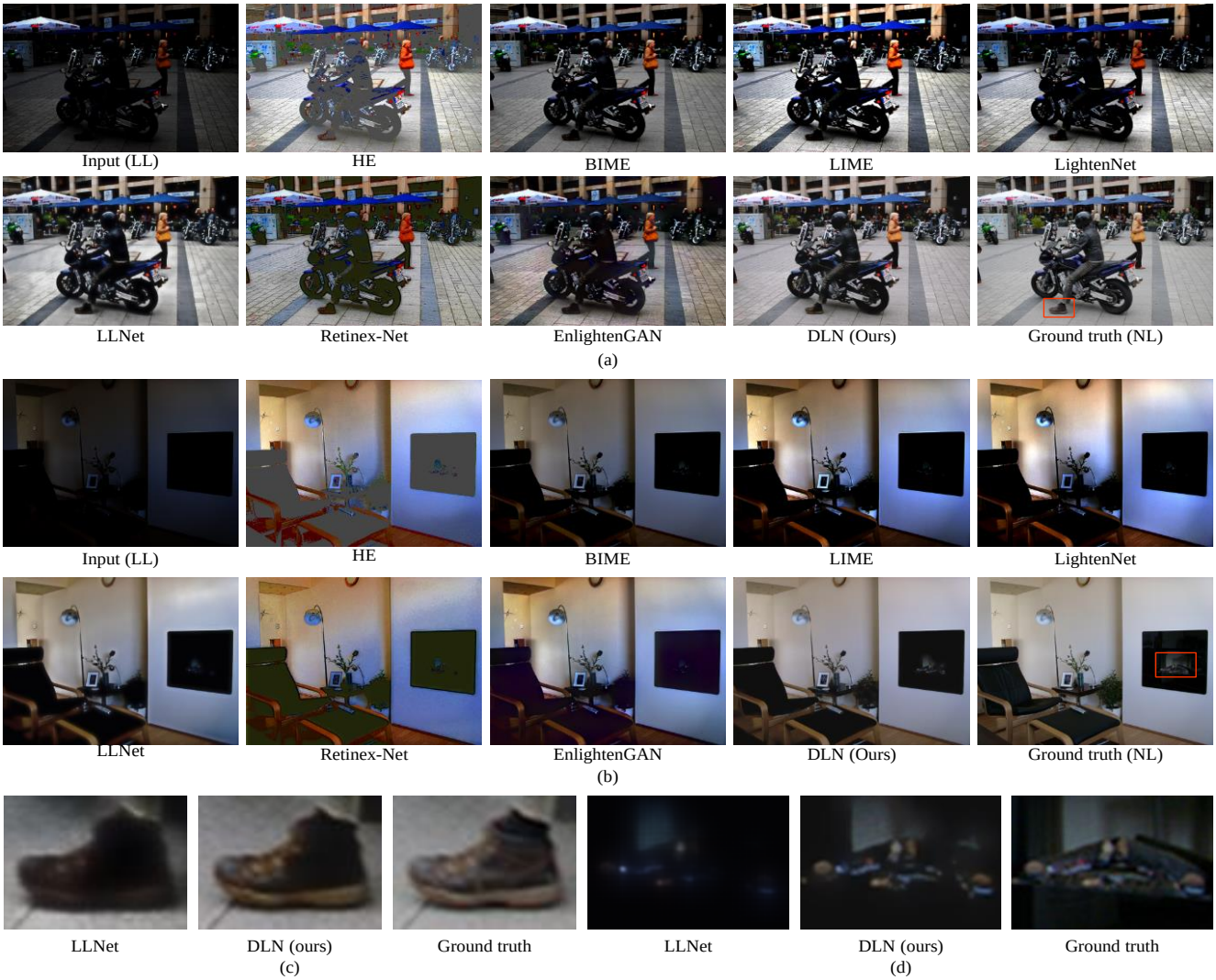

Figure 9. Visual comparison of different algorithms on synthetic dataset (zoom in for a better view)

cannot see the content of the screen from the estimation of the LLNet, while the result from DLN is more recognizable.

C. *Evaluation on the Real Dataset*

LOw-Light (LOL) dataset [20] contains 500 (485 for training and 15 for testing) paired LL-NL images that were captured by a camera with two different exposure times and ISO for each scene. It is the first dataset for the evaluation of low-light enhancement for real scenes. We fine-tuned the model on the LOL training dataset and then evaluated it on the LOL testing dataset.

TABLE VIII. COMPARISON ON REAL DATASET (**RED**: BEST; **BLUE**: THE 2ND BEST, **GREEN**: THE 3RD BEST)

Method	PSNR	SSIM	NIQE
HE	15.467	0.504	9.531
BIMEF	13.875	0.577	7.699
LIME	16.92	0.599	8.795
LightenNet	10.301	0.361	7.422
LLNet	**17.953**	**0.704**	**3.974**
Retinex-Net	16.774	0.559	9.728
EnlightenGAN	**17.483**	**0.658**	**4.889**
DLN (proposed)	**21.946**	**0.807**	**3.656**

Let us compare the DLN with the existing methods, and the results are shown in Table VIII. It can be seen from the table that for both PSNR and SSIM, the proposed DLN approach achieves the best performance with an average PSNR score of 21.946 dB and SSIM score of 0.807, which exceed the second-best approach (LLNet) by 3.993 dB (21.946-17.953) on PSNR and 0.103 (0.807-0.704) on SSIM. It suggests that the proposed DLN approach has excellent lightening ability that achieves the best result of the low-light enhancement among all compared methods. Recently, some metrics [37-40] have been proposed which focus on the visual quality evaluation. NIQE [37] measures the visual quality based on natural scenes statistics. A smaller NIQE score indicates better visual quality. We can see from TABLE VIII that the proposed method obtains the best NIQE score (3.656), which means the predicted NL images are with the best visual quality.

Fig. 10 shows the visualization comparison of different approaches. It can be seen from Fig. 10(a) that the result of our DLN approach significantly improves the visual quality of the input low-light image, which is brighter than other approaches, including BIME, LightenNet, LLNet, and EnlightenGAN. Furthermore, the hue of our DLN is similar to the ground-truth NL images, while other approaches (like HE, LIME, and Retinex-Net, see for example, the color 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.2020.3008396, IEEE Transactions on Image Processing

Figure 10. Visual comparison of different algorithms on real dataset (zoom in for a better view)

the chairs) suffer from serious color shift. Qualitative results show that our DLN, LLNet, and LightenNet achieve the top-3 best performance. Fig. 10(b) shows the visual comparison of these methods. We can see from the figure that all the methods can enhance the low-light image effectively. However, the results of the LLNet and EnlightenGAN include some black stain (see the yellow circle in the figure) that reduces the visual quality. Besides, for the detailed reconstruction, like the color stripe of the diving platform, as shown in the red rectangle of Fig. 10(b), our DLN produces a recognizable result that we can easily distinguish the blue and red stripes.

User study: We have also performed a user study to qualify the visual similarity between the enhancement results and ground truth NL images. In the user-study, the users were requested to evaluate the similarity between the enhancement results and the ground truth ones. The similarity is divided into five levels, where the score 5 means "exactly the same" and 1 denotes "totally different". We used three images chosen from the LOL testing dataset, and all estimations were presented blindly. We invited 85 users to finish our user study. The result is shown in TABLE IX, where Mean Opinion Score (MOS) averages the scores of all users. Retinex-Net obtains the lowest score (2.43), which means the estimation has obvious differences to the ground-truth. The LLNet (2.76) and EnlightenGAN (2.67) obtained similar scores in the experiment, which is consistent with the previous quantitative result in terms of PSNR (LLNet: 17.953 dB, EnlightenGAN: 17.483dB). Although LIME has lower PSNR (16.92 dB), its estimation has slightly better perceptual quality (score of 2.91) than the LLNet (2.76) and EnlightenGAN (2.67). It is clear that the proposed DLN gives the best result (3.12), whose estimations are most similar to the ground truth images.

TABLE IX. USER STUDY

	LIME	LLNet	Retinex-Net	Enlighten GAN	DLN (proposed)
MOS	2.91	2.76	2.43	2.67	**3.12**

V. CONCLUSION

In this paper, we have introduced our proposed Deep Lightening Network (DLN) for low-light image enhancement. Unlike the previous methods that either learn the mapping between the low- and normal-light images directly, or adopt GAN-based method for perception reconstruction, we propose a novel Lightening Back-Projection (LBP) block which learns the differences between the low- and normal-light images iteratively. To strengthen the representation power of the input of the lightening process, we fuse the feature maps with different receptive fields through the Feature Aggregation (FA) block, which is an extension of the squeeze-and-extension structure that investigates both the spatial and channel-wise dependencies among different feature maps. Benefited from the residual estimation of LBP and the rich features of the FA, the proposed DLN gives a better reconstruction of the normal-light condition. Besides, the network works in an end-to-end way, which makes it easy to implement. We have used both objective and subjective evaluations to compare the performance of the proposed DLN with other methods. Extensive results show that our proposed method outperforms other recent state-of-the-art approaches (conventional, CNN-based, and GAN-based methods) in quantitative and qualitative aspects.

In the further work, we can continue to explore more effective CNN structures to improve the performance of the low-light enhancement, and investigate methods for low-light video enhancement. The quality of the simulated LL-NL images is highly related to the performance of the trained model. A good simulation can obviously improve the generalization ability of the enhancement model, which is an interesting research topic to be investigated in the future. Also, our proposed algorithm dramatically improves the visibility of the low-light images, which can be used in various applications. For example, it can be used in a driving assistant system to provide reliable visual aid for a dark and difficult environment.

ACKNOWLEDGMENT

This work was supported in part with a PhD research studentship to Li-Wen Wang by The Hong Kong Polytechnic University.

REFERENCES

[1] Etta D Pisano, Shuquan Zong, Bradley M Hemminger, Marla DeLuca, R Eugene Johnston, Keith Muller, M Patricia Braeuning and Stephen M Pizer, "Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms," *Journal of digital imaging*, vol. 11, no. 4, pp. 193, 1998.

[2] Mohammad Abdullah-Al-Wadud, Md Hasanul Kabir, M Ali Akber Dewan and Oksam Chae, "A dynamic histogram equalization for image contrast enhancement," *IEEE transactions on consumer electronics*, vol. 53, no. 2, pp. 593-600, 2007.

[3] Zia-ur Rahman, Daniel J Jobson and Glenn A Woodell, "Retinex processing for automatic image enhancement," *Journal of electronic imaging*, vol. 13, no. 1, pp. 100-111, 2004.

[4] Jin-Hwan Kim, Jae-Young Sim and Chang-Su Kim, "Single image dehazing based on contrast enhancement," *Proceedings, IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1273-1276, 2011, Prague, Czech Republic.

[5] L. Li, R. Wang, W. Wang and W. Gao, "A low-light image enhancement method for both denoising and contrast enlarging," *Proceedings, IEEE international conference on image processing (ICIP)*, pp. 3730-3734, 2015, Québec, Canada.

[6] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Proceedings, Advances in neural information processing systems*, pp. 1097-1105, 2012.

[7] Spyros Gidaris and Nikos Komodakis, "Object detection via a multi-region & semantic segmentation-aware CNN model," *Proceedings, ICCV*, 2015.

[8] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li and Wan-Chi Siu, "Hierarchical Back Projection Network for Image Super-Resolution," *Proceedings, IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, pp. 0-0, 2019, California, United States.

[9] R. Girshick, "Fast R-CNN," *Proceedings, 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440-1448, 2015.

[10] Matthew D Zeiler and Rob Fergus, "Visualizing and understanding convolutional networks," *Proceedings, European conference on computer vision (ECCV)*, pp. 818-833, 2014, Zurich.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep residual learning for image recognition," *Proceedings, IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 770-778, 2016, Las Vegas, United States.

[12] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke and Andrew Rabinovich, "Going deeper with convolutions," *Proceedings, IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1-9, 2015, Boston, Massachusetts.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten and Kilian Q Weinberger, "Densely connected convolutional networks," *Proceedings, IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4700-4708, 2017, Hawaii, United States.

[14] Chao Dong, Chen Change Loy, Kaiming He and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," *Proceedings, European conference on computer vision (ECCV)*, pp. 184-199, 2014, Zurich, Switzerland.

[15] Chao Dong, Chen Change Loy, Kaiming He and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 38, no. 2, pp. 295-307, 2015.

[16] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao and Lei Zhang, "Toward real-world single image super-resolution: A new benchmark and a new model," *Proceedings, IEEE international conference on computer vision (ICCV)*, pp. 3086-3095, 2019, South Korea.

[17] Muhammad Haris, Gregory Shakhnarovich and Norimichi Ukita, "Deep back-projection networks for super-resolution," *Proceedings, IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1664-1673, 2018, Utah, United States.

[18] Zhi-Song Liu, Wan-Chi Siu and Yui-Lam Chan, "Joint Back Projection and Residual Networks for Efficient Image Super-Resolution," *Proceedings, 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1054-1060, 2018.

[19] Muhammad Haris, Gregory Shakhnarovich and Norimichi Ukita, "Deep back-projection networks for super-resolution," *Proceedings, Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1664-1673, 2018.

[20] Chen Wei, Wenjing Wang, Wenhan Yang and Jiaying Liu, "Deep retinex decomposition for low-light enhancement," *Proceedings, British Machine Vision Conference (BMVC)*, 2018, Newcastle, UK.

[21] Chongyi Li, Jichang Guo, Fatih Porikli and Yanwei Pang, "Lightennet: A convolutional neural network for weakly illuminated image enhancement," *Pattern recognition letters*, vol. 104, pp. 15-22, 2018.

[22] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou and Zhangyang Wang, "EnlightenGAN: Deep Light Enhancement without Paired Supervision," *arXiv preprint arXiv:1906.06972*, 2019.

[23] Martin Arjovsky, Soumith Chintala and Léon Bottou, "Wasserstein gan," *arXiv preprint arXiv:1701.07875*, 2017.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens and Zbigniew Wojna, "Rethinking the inception architecture for computer vision," *Proceedings, IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2818-2826, 2016, Las Vegas, United States.

[25] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *Proceedings, AAAI Conference on Artificial Intelligence*, 2017, California, USA.

[26] Jie Hu, Li Shen and Gang Sun, "Squeeze-and-excitation networks," *Proceedings, IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7132-7141, 2018, Utah, United States.

[27] Feifan Lv, Feng Lu, Jianhua Wu and Chongsoon Lim, "MBLLEN: Low-light image/video enhancement using CNNs," *Proceedings, British Machine Vision Conference (BMVC)*, 2018, England.

[28] Zhou Wang, Alan C Bovik, Hamid R Sheikh and Eero P Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing (TIP)*, vol. 13, no. 4, pp. 600-612, 2004.

[29] gregjohnso. "Github: pytorch ssim," Access date: 01 Nov, 2019; Retrieved from https://github.com/Po-Hsun-Su/pytorch-ssim/tree/master/pytorch_ssim.

[30] Zhenqiang Ying, Ge Li and Wen Gao, "A bio-inspired multi-exposure fusion framework for low-light image enhancement," *arXiv preprint arXiv:1711.00591*, 2017.

[31] Xiaojie Guo, Yu Li and Haibin Ling, "LIME: Low-light image enhancement via illumination map estimation," *IEEE transactions on image processing (TIP)*, vol. 26, no. 2, pp. 982-993, 2016.

[32] Kin Gwn Lore, Adedotun Akintayo and Soumik Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern recognition*, vol. 61, pp. 650-662, 2017.

[33] Feifan Lv and Feng Lu, "Attention-guided Low-light Image Enhancement," *arXiv preprint arXiv:1908.00682*, 2019.

[34] Alex Clark and Contributors. "Pillow: a python imaging library," Access date: 24 Sep, 2019; Retrieved from https://python-pillow.org.

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *Proceedings, IEEE international conference on computer vision (ICCV)*, pp. 1026-1034, 2015, Las Condes, Chile.

[36] Hang Zhao, Orazio Gallo, Iuri Frosio and Jan Kautz, "Loss functions for image restoration with neural networks," *IEEE transactions on computational imaging*, vol. 3, no. 1, pp. 47-57, 2016.

[37] Anish Mittal, Rajiv Soundararajan and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE signal processing letters*, vol. 20, no. 3, pp. 209-212, 2012.

[38] Anish Mittal, Anush Krishna Moorthy and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE transactions on image processing (TIP)*, vol. 21, no. 12, pp. 4695-4708, 2012.

[39] Lin Zhang, Lei Zhang, Xuanqin Mou and David Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE transactions on image processing (TIP)*, vol. 20, no. 8, pp. 2378-2386, 2011.

[40] Si Lu, "No-reference Image Denoising Quality Assessment," *Proceedings, Science and information conference*, pp. 416-433, 2019.

Li-Wen Wang received his Bachelor of Engineering degree, in 2016, from Shandong University. Subsequently, he obtained his Master of Science (MSc) degree with distinction from The Hong Kong Polytechnic University, where he is now a second-year Ph.D. student under the supervision of Professor Wan-Chi Siu and Dr. Daniel P. K. Lun. His research interests include deep learning techniques, image and video processing, object detection and tracking, and autonomous driving.

Zhi-Song Liu received the MSc degree in electronic engineering, in 2015, from The Hong Kong Polytechnic University, Hong Kong, where he is currently working toward the PhD degree under the supervision of Prof. Wan-Chi Siu and Dr. Yui-Lam Chan. His research interests include deep learning techniques, image and video signal processing, image and video super-resolution

Wan-Chi Siu (S'77-M'77-SM'90-F'12-Life-F'16) received the MPhil and PhD degrees from The Chinese University of Hong Kong in 1977 and Imperial College London in 1984. He is Life-Fellow of IEEE and Fellow of IET, and Immediate-Past President (2019-2020) of APSIPA (Asia-Pacific Signal and Information Processing Association). Prof. Siu is now Emeritus Professor, and was Chair Professor, Founding Director of Signal Processing Research Centre, Head of Electronic and Information Engineering Department and Dean of Engineering Faculty of The Hong Kong Polytechnic University. He is an expert in DSP, transforms, fast algorithms, machine learning, and conventional and deep learning approaches for super-resolution imaging, 2D and 3D video coding, object recognition and tracking. He has published 500 research papers (over 200 appeared in international journal papers), and edited three books. He has also 9 recent patents granted. Prof. Siu was an independent non-executive director (2000-2015) of a publicly-listed video surveillance company and convenor of the First Engineering/IT Panel of the RAE(1992/93) in Hong Kong. He is an outstanding scholar, with many awards, including the Best Teacher Award, the Best Faculty Researcher Award (twice) and IEEE Third Millennium Medal (2000). Prof. Siu has been Guest Editor/Subject Editor/AE for IEEE Transactions on Circuits and System II, Image Processing, Circuit & System for Video Technology, and Electronics Letters, and organized very successfully over 20 international conferences including IEEE society-sponsored flagship conferences, such as TPC Chair of ISCAS1997 and General Chair of ICASSP2003 and General Chair of ICIP2010. He was Vice-President, Chair of Conference Board and Core Member of Board of Governors (2012-2014) of the IEEE Signal Processing Society, and is now a member of the IEEE Educational Activities Board, IEEE Fourier Award for Signal Processing Committee and some other IEEE committees.

Daniel P. K. Lun (SM'12) received the B.Sc. degree (Hons.) from the University of Essex, U.K., in 1988, and the Ph.D. degree from The Hong Kong Polytechnic University in 1991. He is currently an Associate Professor with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University. He is active in research. He has published over 130 international journal and conference papers. His research interests include signal and image enhancement, sparse representation and applications, and 3D data acquisition. He is a member of the DSP and Visual Signal Processing and Communications Technical Committee, and the IEEE Circuits and Systems Society. He is a Chartered Engineer, a fellow of IET, and a Corporate Member of HKIE. He was a General Co-Chair, a Technical Co-Chair, and an Organizing Committee Member of a number of international conferences, including ICASSP 2003, ICIP 2010, and ICME 2017. He was the Chairman of the IEEE Hong Kong Chapter of Signal Processing from 1999 to 2000. He is currently an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.