Neurocomputing 518 (2023) 332–343

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Low-light image enhancement with knowledge distillation

Ziwen Li [a], Yuehuan Wang [a,*], Jinpu Zhang [a]

[a] National Key Lab of Science and Technology on Multi-spectral Information Processing, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 21 January 2022
Revised 1 September 2022
Accepted 30 October 2022
Available online 11 November 2022
Communicated by Zidong Wang

Keywords:
Low light image enhancement
Knowledge distillation
Deep learning

ABSTRACT

Low-light image enhancement studies how to improve the quality of images captured under poor lighting conditions, which is of real-world importance. Currently, convolutional neural network (CNN)-based methods with state-of-the-art performance have become the mainstream of research. However, most CNN-based methods improve the performance of the algorithm by increasing the width and depth of the neural network, which requires large computing device resources.

In this paper, we propose a knowledge distillation method for low light image enhancement. The proposed method uses a teacher-student framework in which the teacher network tries to transfer the rich knowledge to the student network. The student network learns the knowledge of image enhancement under the supervision of ground truth images and under the guidance of the teacher network simultaneously. Knowledge transfer between the teacher-student network is accomplished by distillation loss based on attention maps. We designed a gradient-guided low-light image enhancement network that can be divided into an enhancement branch and a gradient branch, where the enhancement branch is learned under the guidance of the gradient branch to better preserve structural information. The teacher and student networks use a similar structure, but they have different model sizes. The teacher network has more parameters and more powerful learning capabilities than the student network. With the help of knowledge distillation, our approach can improve the performance of the student network without increasing the computational burden during the testing phase. The qualitative and quantitative experimental results demonstrate the superiority of our method compared to the state-of-the-art methods.

1. Introduction

Low-light image enhancement has been studied for many years and has important applications in nighttime video surveillance, unmanned vehicles and photo retouching software. In low-light environments, insufficient light causes the captured images to fail to show image details clearly, thus affecting the visual performance of the images. Meanwhile, low-light images can affect the performance of high-level vision tasks, such as face recognition [1,2] and scene segmentation [3,4].

Traditional methods for low-light image enhancement mainly include histogram equalization-based methods and Retinex-based methods. The histogram equalization-based method [5–9] adjusts the overall brightness of the image by redistributing the overall distribution of pixels with different brightness. And the Retinex-based method [10–18] decomposes the low-light image into an illumination map and a reflection map, and adjusts the illumination map or uses the reflection map directly as the enhancement result. Traditional prior-based methods have been able to achieve good performance. However, the traditional prior-based approach cannot handle a wide range of lighting situations, and the generated results often tend to blur details and suffer from uneven illumination.

Due to the powerful feature representation capability of deep learning, data-driven algorithms have become the mainstream of computer vision. Most current low-light image enhancement methods are based on convolutional neural networks, which fit the mapping relationship from low-light images to normal light images from a large amount of data through a finely designed CNN structure. LLNet [19] is the first CNN-based method that uses a sparse denoising autoencoder to simultaneously learn low-light image enhancement and denoising, opening the door to deep learning methods for image enhancement. Subsequently, many low-light image enhancement methods based on deep learning have been proposed, such as RetinexNet [20], MBLLEN [21], KinD [22] and EnlightenGAN [23], all of which have achieved great

* Corresponding author.
E-mail addresses: D201980722@hust.edu.cn (Z. Li), yuehwang@hust.edu.cn (Y. Wang).

https://doi.org/10.1016/j.neucom.2022.10.083
0925-2312/

progress and impressive visual results. However, most algorithms increase the number of parameters and computation while improving the performance of the algorithm, which requires larger memory space and computing device resources, which is impractical for the real-world applications of the algorithm, such as deployment to mobile platforms and edge devices.

Knowledge distillation is an effective model compression method that can improve the performance of a network without modifying the network structure. Knowledge distillation [24] usually takes a teacher-student framework, where the teacher network uses a complex network and the student network uses a lightweight network, and the competent teacher network transfers the learned knowledge to the student network, thus improving the performance of the student network. Zhang et al. [25] proposed a self-distillation framework, Zagoruyko et al. [26] proposed an attention map-based distillation learning method, and Liu et al. [27,28] proposed pairwise distillation based on similar maps and holistic distillation based on adversarial generative networks. However, most methods have been applied only to high-level vision tasks, such as image classification and semantic segmentation, and few studies have explored the use of knowledge distillation methods to enhance low-light images.

To address the above issues, in this work, we propose a knowledge distillation method for low-light image enhancement. We use a teacher-student framework to transfer knowledge, where the teacher network has more parameters and therefore a stronger function fitting capability, while the student network uses a compressed network structure with a smaller number of parameters. Under the supervision of ground truth images, the teacher network can learn abundant knowledge and store it in the feature maps. The student network not only uses ground truth images as supervision, but also receives the guidance of the teacher network for learning. Attention map-based distillation loss was used to accomplish knowledge transfer between the teacher and student networks. We designed a gradient-guided low-light image enhancement network for both teacher and student networks. The designed network can be divided into two branches, the enhancement branch and the gradient branch. The gradient branch is used to extract high frequency structural information, while the enhancement branch is guided by the gradient branch to learn to preserve more image details. Both the teacher network and the student network use a similar gradient-guided network, but the depth and width of the network are not consistent. The teacher network has a deeper and wider network with more number of parameters and is therefore more powerful than the student network. With the help of knowledge distillation, our approach can improve the upper bound on the performance of student networks without modifying the network structure and without increasing the computational cost during testing. Our approach achieves a trade-off between performance and number of parameters. The experimental results can demonstrate that our method surpasses the existing state-of-the-art methods in a qualitative and quantitative manner.

The main contributions of this work can be summarized as follows.

- We propose a knowledge distillation method for low light image enhancement. The teacher-student network framework is introduced to transfer knowledge from the teacher to the student. With the guidance of the teacher network, our approach can improve the performance of the student network without increasing the computational burden during the testing phase.
- We design a gradient-guided low-light enhancement network that contains two branches, the enhancement branch and the gradient branch, where the enhancement branch learns to preserve more image details under the guidance of the gradient branch.
- We design an attention map based distillation loss to transfer knowledge. Reconstruction loss, SSIM loss, perceptual loss, gradient loss and distillation loss are combined together to complete the network training.
- Experimental results demonstrate that our method outperforms the existing state-of-the-art methods in a qualitative and quantitative manner. Moreover, our approach achieves a balance between performance and computational complexity.

2. Related works

2.1. Low-light image enhancement

2.1.1. Conventional methods

Low-light image enhancement has been studied for many years, mainly based on histogram equalization and Retinex models. Histogram equalization is a classical image enhancement method that stretches the distribution of the histogram so that the brightness can be evenly distributed over the entire range. Some improved methods based on histogram equalization have been proposed by the researchers. Zuiderveld et al. [5] proposes to perform histogram equalization on local areas instead of the whole image and limit the upper limit of contrast amplification to mitigate the problem of noise amplification. Stark et al. [6] proposes a cumulative function formulation to generate a mapping of pixel gray levels from a local histogram. Lee et al. [7] observed that grayscale differences in high-frequency regions should be more pronounced and used a two-dimensional histogram in a tree-like hierarchy to represent grayscale differences. Coltuc et al. [8] transforms the problem into K-dimensional space to obtain a reversible cumulative distribution function, and derives a strict image pixel order.

Based on color constancy, retinex theory [10] can decompose the original image into an illumination map and a reflection map, in which the reflection map represents the essential information of the object, and the illumination map represents the light shining on the object. SSR [11] proposes to solve the illumination map and reflection map in the logarithmic domain using Gaussian surround functions and use the reflection map as the enhancement result. MSR [12] is a multi-scale Retinex model that uses Gaussian filters of different scales to process the original image. MSRCR [12] proposed the color recovery function to adjust the ratio of the three color components to avoid the color distortion problem in MSR. Guo et al. [13,14] uses the maximum pixel value as the initial illumination map and combines the structural smoothing prior information to optimize the illumination map using the augmented Lagrange multiplier method. Wang et al. [15] designs bright-pass filters for non-uniform images to perform Retinex decomposition, and uses a double logarithmic transform to try to preserve details and naturalness. Li et al. [16] proposes to add a noise component to the original Retinex model to estimate the noise map with constraints on the segmental smoothness of the illumination map and the consistency of the gradient of the reflectance map. Fu et al. [18] proposes to use a weighted variational model to better estimate the illumination and reflectance maps. Fu et al. [17] uses a morphological closed-loop operation to complete the decomposition of the illumination and reflection maps, and then uses a weighted fusion method to fuse the two illumination maps obtained using curve adjustment and histogram equalization. However, the methods based on histogram equalization and Retinex model tend to lose the image details and blur edge information, with limited recovery.

2.1.2. Deep learning methods

Due to the rapid development of deep learning in computer vision, deep learning techniques have been successfully applied to the field of low-light image enhancement and have occupied a

mainstream research position. Lore et al. [19] proposed to use an autoencoder structure for simultaneous low-light image enhancement and denoising, which is the first application of deep learning to low-light image enhancement. Lv et al. [21] proposed a multi-branch enhancement network to extract features at different levels and fusion to get the output image. Wang et al. [29] proposed the global illumination perception and detail preservation network for image enhancement. Wei et al. [20], based on Retinx theory, proposed the network to decompose the image into illumination and reflection maps. In addition, a low-light image dataset LOL [20] is also proposed. Wang et al. [30] extracts global and local features at low resolution, then uses bilateral grid based upsampling to get a full-resolution illumination map. Zhang et al. [22] first decomposes the original image into a illumination map and a reflection map using a decomposition network, and then adjusts the light map and reflection map using an adjustment network and a reflection map recovery network, respectively. Wang et al. [31] proposed the light back-projection module and built the Deep Lightening Network in a cascade manner, and finally introduced an adjustable light control factor parameter to enhance the low-light image. Zhang et al. [32] proposed to use optical flow estimation method to simulate video from a single image to ensure temporal consistency of low-light video enhancement. Liu et al. [33] proposed a neural network architecture search method, based on Retinex theory prior knowledge, to automatically search for efficient network frames from a predefined search space. Yang et al. [34] proposed a GAN-based domain adaptation mechanism to learn from both paired and unpaired data. Jiang et al. [35] proposed the Retinex-based self-regularized method and recovered in HSV space to avoid color bias. Jiang et al. [23] proposes global and local discriminators to enhance low-light images using generative adversarial networks for unpaired learning. Guo et al. [36] proposes a method based on high-order curve adjustment without ground truth images, using several unsupervised losses to drive the training of the network. Yang et al. [37] proposes a semi-supervised approach to enhance images, with supervised learning to restore image fidelity and unsupervised learning to improve the perceptual quality of images.

2.2. Knowledge distillation

Knowledge distillation is an effective model compression technique that can effectively reduce the size of the network model. Hinton et al. [24] first introduced the concept of knowledge distillation and designed a teacher-student framework in which the performance of the student network was improved by migrating the soft label distribution. Chen et al. [38] proposes a technique based on function-preserving transformations to migrate an already trained network to another neural network, which can accelerate the training of deeper and wider large networks. Lin et al. [39] proposes a holistic framework for knowledge distillation, a scheme that eliminates the redundancy between fully connected and convolutional layers by using a low-rank decomposition strategy. Zhang et al. [25] proposed self-distillation framework in which deep features supervise the learning of shallow features. Zagoruyko et al. [26] proposed a distillation method based on attentional maps, by using attentional features as knowledge. Liu et al. [27,28] proposed pairwise distillation based on similar maps and holistic distillation based on adversarial generative networks. Li et al. [40] applies knowledge distillation to object detection, uses MSE to optimize the similarity of features between teacher and student, and uses a transformation layer to keep teacher and student network features dimensionally consistent. For the object detection task, Zhang et al. [41] proposed attentional distillation for learning salient features of foreground targets and using nonlocal distillation to learn correlations between different pixels. Gao et al. [42], He et al. [43], Lee et al. [44] proposed to apply knowledge distillation to single image super-resolution. Hong et al. [45] proposed to distill the knowledge of reconstructing clear images to the student network. However, most knowledge distillation methods are based on high-level vision tasks, and few studies have explored how knowledge distillation can be used for low-light image enhancement.

3. Method

In this section, we first describe an overview of the knowledge distillation framework, including the teacher-student framework and their learning patterns. Then, we describe the loss functions used for optimization, including reconstruction loss, perceptual loss, SSIM loss, gradient loss, and distillation loss. Finally, we describe the gradient-guided network structure.

3.1. Overview

Our knowledge distillation framework uses a teacher-student framework in which the teacher guides the students for better

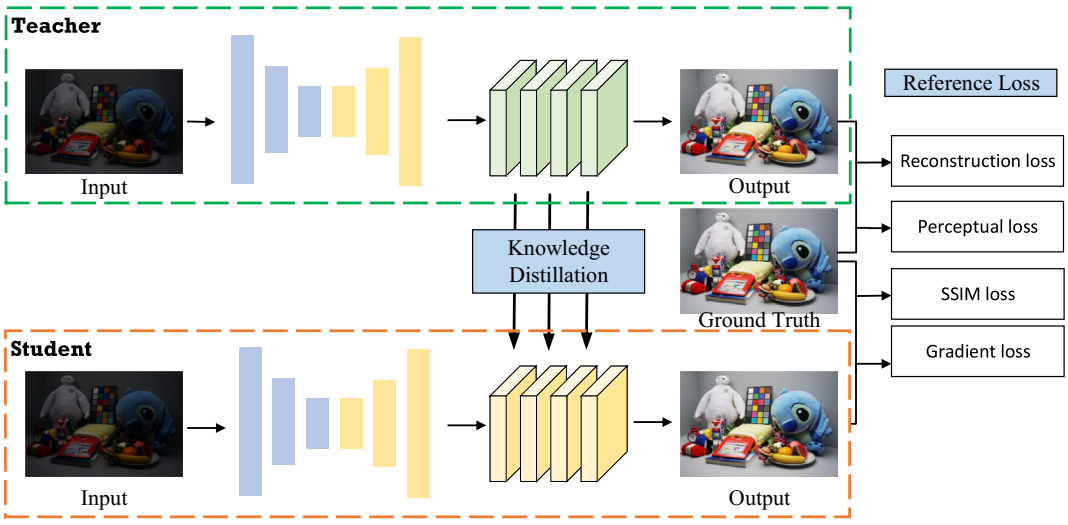


Fig. 1. The overall framework of knowledge distillation for low-light image enhancement.

learning. Large networks can learn rich knowledge of low-light image enhancement, which is stored in trained convergent network features. In deep learning, networks with a large number of parameters will have a stronger feature representation, while smaller networks will have difficulty outperforming larger networks due to fewer network parameters. The motivation of knowledge distillation is to distill the knowledge learned in large networks and transfer this knowledge to small networks with a small number of parameters, thus improving the learning ability of small networks. In general, large networks are referred to as teacher networks, denoted by G_T, while small networks are referred to as student networks, denoted by G_S.

As shown in the Fig. 1, the teacher network can learn rich knowledge of low-light image enhancement under the supervision of the reference image, and its network parameters can be optimized with reference loss L_{ref} based on the reference image supervision. While the student network learns under the supervision of both the reference image and the teacher network, and its parameter learning process is optimized using both the reference loss L_{ref} and the distillation loss L_{dis}. The specific reference losses will be described later. The learning process can be expressed as

$$Teacher: \ \arg\min_{w} L_{ref}(G_T(I,w), GT) \tag{1}$$

$$Student: \ \arg\min_{w} L_{ref}(G_S(I,w), GT) + \beta L_{dis}(F_T, F_S), \tag{2}$$

where I and GT denote the input image and the ground truth image, $G_T(I,w)$ and $G_S(I,w)$ denote the generated images of the teacher network and the student network. In the distillation loss, F_T and F_S denote the feature maps from the teacher network and the student network, respectively. The parameter β is a trade-off parameter between the reference loss and the distillation loss.

We use a two-stage training process to train our knowledge distillation network, as shown in the Algorithm 1. In the first stage, the teacher network is first trained to reach the convergence state. In the second stage, we fix the trained teacher network model and use it to train the student network. The two-stage training mode allows reusing the trained teacher model and extracting the best teacher knowledge at the beginning.

Algorithm 1 The training process of the proposed method

STAGE 1: Training teacher network.
INPUTS: Initialized teacher network G_T, pairwise training data I and GT.
$\arg\min_w L_{ref}(G_T(I,w), GT))$
STAGE 2: Training student network.
INPUTS: Already trained teacher network G_T, initialized student network G_S, pairwise training data I and GT.
$\arg\min_w L_{ref}(G_S(I,w), GT) + \beta L_{dis}(F_T, F_S)$

3.2. Loss function

3.2.1. Loss function for teacher network

The training of the teacher network is optimized by reference loss based on the ground truth image. Following previous work [22,37], we combine L1 loss L_1, perceptual loss L_{per}, SSIM loss L_{ssim} and gradient loss L_{grad} to train teacher networks. The overall objective function of the network training can be defined as

$$L_T = L_1 + \sigma L_{per} + \alpha L_{ssim} + \gamma L_{grad}, \tag{3}$$

where σ, α, γ denote the trade-offs of different loss functions.

The L1 loss is a common pixel-level image reconstruction loss that reduces the pixel-level differences, and the formula can be expressed as

$$L_1 = \|G_T(I), GT\|_1. \tag{4}$$

The perceptual loss can be used to improve the perceptual quality of the generated images. The perceptual features of the image are extracted from the pre-trained network, and the L1 metric is used to calculate the difference with the ground truth image, which can be defined by the formula

$$L_{per} = \|\Phi(G_T(I)), \Phi(GT)\|_1, \tag{5}$$

where Φ denotes the extraction of features from the VGG19 network.

SSIM is a commonly used image quality assessment metric by measuring the similarity of the generated image to the reference image in terms of brightness, contrast and structure. Here, we use it to constrain the training of the network, which can be expressed as

$$L_{ssim} = 1 - SSIM(G_T(I), GT)). \tag{6}$$

Gradient loss is often used to preserve high frequency texture details of an image and to avoid blurring of the image. Here the gradient is calculated using Sobel and the distance between gradient maps is calculated using L2 norm distance.

$$L_{grad} = \|\nabla(G_T(I)), \nabla(GT)\|_2. \tag{7}$$

3.2.2. Loss function for student network

The student network is not only supervised by the ground truth images, but also needs to be guided by the teacher network. The loss functions for student network training include L1 loss L_1, perceptual loss L_{per}, SSIM loss L_{ssim}, gradient loss L_{grad} and distillation loss L_{dis}. The overall loss of the student network can be expressed as

$$L_S = L_1 + \sigma L_{per} + \alpha L_{ssim} + \gamma L_{grad} + \beta L_{dis}, \tag{8}$$

where the reference loss are consistent with the loss of the teacher network and βdenotes the weight parameter of the distillation loss.

The purpose of distillation loss is to make the student network learn the teacher's knowledge and improve the performance of the student network. Inspired by the attention transfer [26], we design a distillation loss based on the attention map to force the attention pattern of the student network to be as close as possible to that of the teacher network. Let the extracted network features be F and their dimensions be $C \times H \times W$, where C, H, and W are the channel, height, and width of the features F, respectively. In order to transform the feature map of $C \times H \times W$ into an attention map of $H \times W$, we can compress it to a single channel dimension by summing the values of each channel at each spatial location, which can be expressed using the following equation

$$Q = \frac{1}{\|F\|_2^2} \sum_{i=1}^{C} |F_i|, \tag{9}$$

where Q denotes the generated normalized attention map and F_i denotes the i-th channel feature of F.

We select the feature maps of the last layer of the teacher network and the student network as the knowledge to be transferred. By summing the values of each spatial location over the channel dimension, we can obtain the spatial attention maps of teachers and students, respectively. When training the student network, the student's attention map will be close to the teacher's in order to transfer the knowledge to the student. We use the mean squared error to measure the distance between the two attention maps. The distillation loss can be defined as

$$L_{dis} = \|Q_T - Q_S\|_2, \tag{10}$$

where Q_T denotes the teacher's attention map and Q_S denotes the student's attention map.

3.3. Network architecture

To balance the network performance and the number of parameters, we design a gradient-guided low-light image enhancement network. As shown in the Fig. 2, our gradient-guided network can be divided into an enhancement branch and a gradient branch, where the gradient branch extracts gradient features while the enhancement branch learns to enhance low-light images under the guidance of the gradient branch. The guidance of the gradient branch can better help the network retain structural information.

The enhancement branch is used as the main part of the gradient-guided network structure, using an autoencoder-like design. First, we use stride-2 convolution for downsampling, which allows the network to compute on low resolution feature maps, reducing the computational effort of the network. Then, the network goes through several successive feature transformation modules, which are mainly designed to learn complex feature mappings. We introduce the FAB (Feature Attention Block) in FFA-Net [46] as our feature transformation module to model the spatial correlation and channel weights of the images, which consists of a residual module, pixel attention and channel attention. Finally, the stride-2 transposed convolution is used to upsample the features to recover them to the resolution size of the original image. We discard all pooling and normalization layers. This is because pooling layers will inevitably cause loss of useful information, while normalization layers will cause contrast bias, which could potentially harm the performance of the network.

Image gradients reveal the overall structural information of an image, including the edges and corner points of the image. To explicitly extract the structural details of an image, we design a gradient branch based on the structural prior and integrate it into the enhancement branch to guide the generation of high-quality images. We use the Sobel operator for gradient map computation, with the following formula defined:

$$G_x(x,y) = Sobel_x * I(x,y), \tag{11}$$

$$G_y(x,y) = Sobel_y * I(x,y), \tag{12}$$

$$g(x,y) = \sqrt{G_x^2 + G_y^2}, \tag{13}$$

where g(x,y) denotes the gradient amplitude at position (x,y).

The gradient branch processes the gradient image instead of the original image. As with the enhancement branch, we first downsample the gradient image using a convolution operation with a step size of 2. Then the feature transform is performed using the same number of FABs as in the enhancement branch. The size and number of these gradient features are kept the same as the enhancement features of the enhancement branch for the subsequent fusion of the two branches. Finally, we integrate the gradient features into the enhancement branch to help the network retain better details. Once the network generates a gradient feature, we use summation to fuse the gradient features with the corresponding enhancement features, and then generate the next gradient feature and enhancement features. In this way, all the structural information extracted from the gradient branches can guide the enhancement branches and improve the quality of the enhanced images.

The teacher network and the student network have exactly the same network framework, but different dimensions, such as the number of FABs, the number of feature map channels, etc. Therefore, the number of parameters for the teacher network is more than that for the student network. Specifically, for the teacher network, the number of FABs is 6, and the number of intermediate feature map channels is 96, while the number of FABs for the student network is 3 and the number of intermediate feature map channels is reduced to 48.

4. Experiments

4.1. Experiment setup

Datasets and Evaluation. We use the LOL [20], MIT-Adobe FiveK [47], and SID [48] datasets to evaluate the performance of the proposed method. The LOL dataset is the first captured low-light image dataset containing 500 pairs of low-light and normal-light images, in which 485 pairs are used for training and the remaining 15 pairs are used for testing. The MIT-Adobe FiveK dataset contains a total of 5000 images, of which the expert C results are usually used as ground truth images. For the division of the MIT-Adobe FiveK dataset, 4500 pairs of images were selected for training, while the remaining 500 pairs were used for testing. The SID dataset was obtained by changing the exposure time, where the short exposure images have very low brightness and severe noise, while the long exposure images were used as the corresponding ground truth images. We choose images taken by the Sony camera for

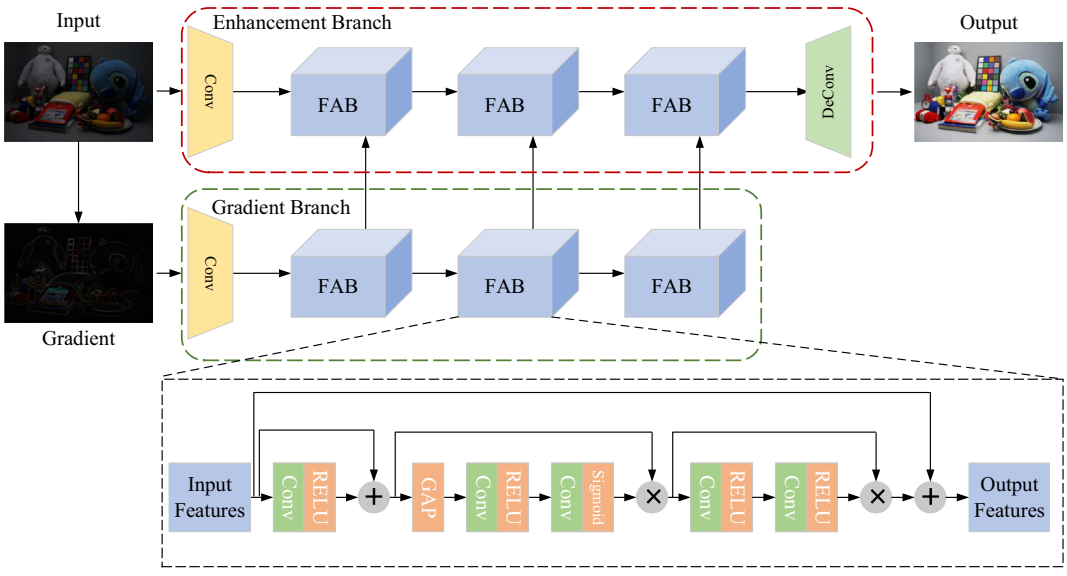


Fig. 2. The architecture of a gradient-guided low-light image enhancement network.

evaluation, where the data format is Bayer mode. We use the rawpy toolkit to convert Raw images to the corresponding RGB images. The training and test sets are divided according to the settings of the [48].

We compared with state-of-the-art methods, including LIME [13], EnlightenGAN [23], ZeroDCE [36], MBLLEN [21], RetinexNet [20], WVM [18], SSIENet [49], KinD [22] and DRBN [37]. We obtain numerical results by running the provided public code or from related papers. We use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) for quantitative evaluation for enhanced results, where larger PSNR and SSIM values are better, while smaller LPIPS values are better.

Implementation Details. All experiments were implemented using the Pytorch framework in the Python environment and run on a GeForce GTX TITAN V GPU device. The input images were randomly cropped into 256×256 patches with horizontal and vertical flips for data augmentation. The Adam optimizer was used to update the network parameters, where the momentum parameters β_1 and β_2 were set to 0.9 and 0.999, respectively. Our student and teacher models were trained for 50 K and 100 K iterations, respectively. The initial learning rate of the student network was set to 0.0002, and the learning rate was halved every 25 K iterations. We first warm up the student network for 10 K iterations using only the reference loss, and then introduce distillation loss to complete the training. Empirically, the loss weights σ, α, γ and β were set to 0.1, 1, 1 and 1, respectively.

4.2. Comparison with state-of-the-arts

4.2.1. Quantitative evaluation

We first performed a quantitative evaluation on these datasets with state-of-the-art methods. The numerical results of these methods are shown in Table 1, Table 2 and Table 3, where the red color indicates the best results and the blue color indicates the second best results. The results of the Teacher model are recorded in the table, but do not participate in the color representation. From the Table 1, we can see that on the LOL dataset, our method achieves the best PSNR value, and from the second and third columns of the table, our method ranks second in the SSIM and LPIPS metrics, lower than DRBN [37], but higher than the other methods, which indicates that our method can well preserve structural details and improve the visual perceptual quality of images. In addition, our method outperforms other methods in the MIT-Adobe FiveK and SID datasets. It is worth noting that both the number of parameters and the computational effort of our method

Table 1
Quantitative comparison results of different methods on the LOL dataset in terms of PSNR, SSIM and LPIPS metrics. The best and second best numerical results are highlighted in red and blue, respectively.

	PSNR↑	SSIM↑	LPIPS↓
LIME	17.17	0.7579	0.1298
WVM	18.44	0.7733	0.1467
MBLLEN	19.08	0.7508	0.1987
RetinexNet	12.50	0.6383	0.2802
SSIENet	9.55	0.6051	0.2760
ZeroDCE	15.85	0.7121	0.2035
EnlightenGAN	16.09	0.7732	0.1464
KinD	16.02	0.7720	0.1542
DRBN	16.81	0.7732	0.1686
Teacher	25.47	0.9134	0.0454
Ours	24.67	0.8911	0.0813

Table 2
Quantitative comparison results of different methods on the MIT-Adobe FiveK dataset in terms of PSNR, SSIM and LPIPS metrics. The best and second best numerical results are highlighted in red and blue, respectively.

	PSNR↑	SSIM↑	LPIPS↓
LIME	17.17	0.7579	0.1298
WVM	18.44	0.7733	0.1467
MBLLEN	19.08	0.7508	0.1987
RetinexNet	12.50	0.6383	0.2802
SSIENet	9.55	0.6051	0.2760
ZeroDCE	15.85	0.7121	0.2035
EnlightenGAN	16.09	0.7732	0.1464
KinD	16.02	0.7720	0.1542
DRBN	16.81	0.7732	0.1686
Teacher	25.47	0.9134	0.0454
Ours	24.67	0.8911	0.0813

Table 3
Quantitative comparison results of different methods on the SID dataset in terms of PSNR and SSIM metrics. The best and second best numerical results are highlighted in red and blue, respectively.

	PSNR	SSIM
LIME	15.52	0.3011
WVM	11.95	0.0382
MBLLEN	15.75	0.4175
RetinexNet	16.49	0.2704
SSIENet	17.24	0.4595
ZeroDCE	15.54	0.2173
EnlightenGAN	17.23	0.5430
KinD	18.02	0.5830
DRBN	19.02	0.5770
Teacher	22.47	0.6681
Ours	21.34	0.6499

(a)

(b)

Fig. 3. Qualitative visual comparison results of the different methods on the LOL dataset.

are relatively smaller compared to most other methods. Overall, our approach achieves competitive performance with smaller computational burden compared to state-of-the-art methods.

4.2.2. Qualitative evaluation

Then we conducted a qualitative comparison between our method and other methods. Fig. 3 shows the visual results of the different methods on the LOL dataset. As shown in Fig. 3, WVM and ZeroDCE lack sufficient brightness, resulting in underexposed images. RetinexNet produces some unwanted artifacts that can be found in the rectangular regions. LIME and SSIENet enhance the images but also amplify the noise (e.g., the two rectangular in Fig. 3b regions). MBLLEN over-smoothes the image, thus blurring the edge detail information. EnlightenGAN suffers from slight

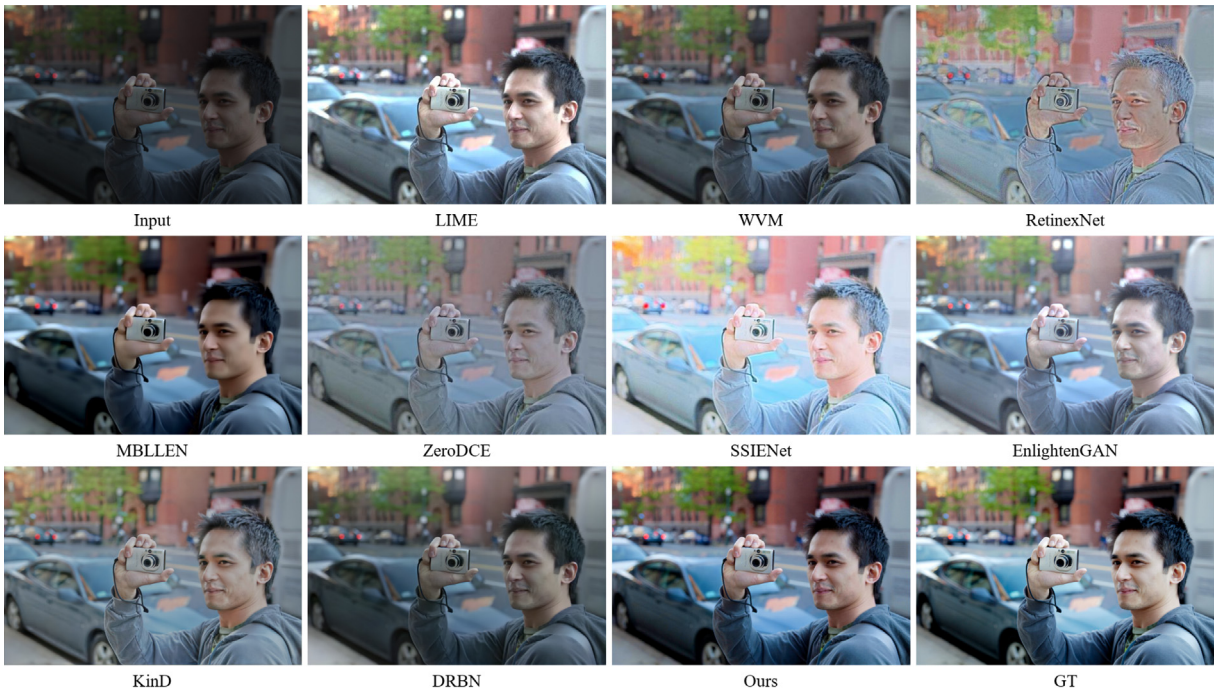


Fig. 4. Qualitative visual comparison results of the different methods on the MIT-Adobe FiveK dataset.

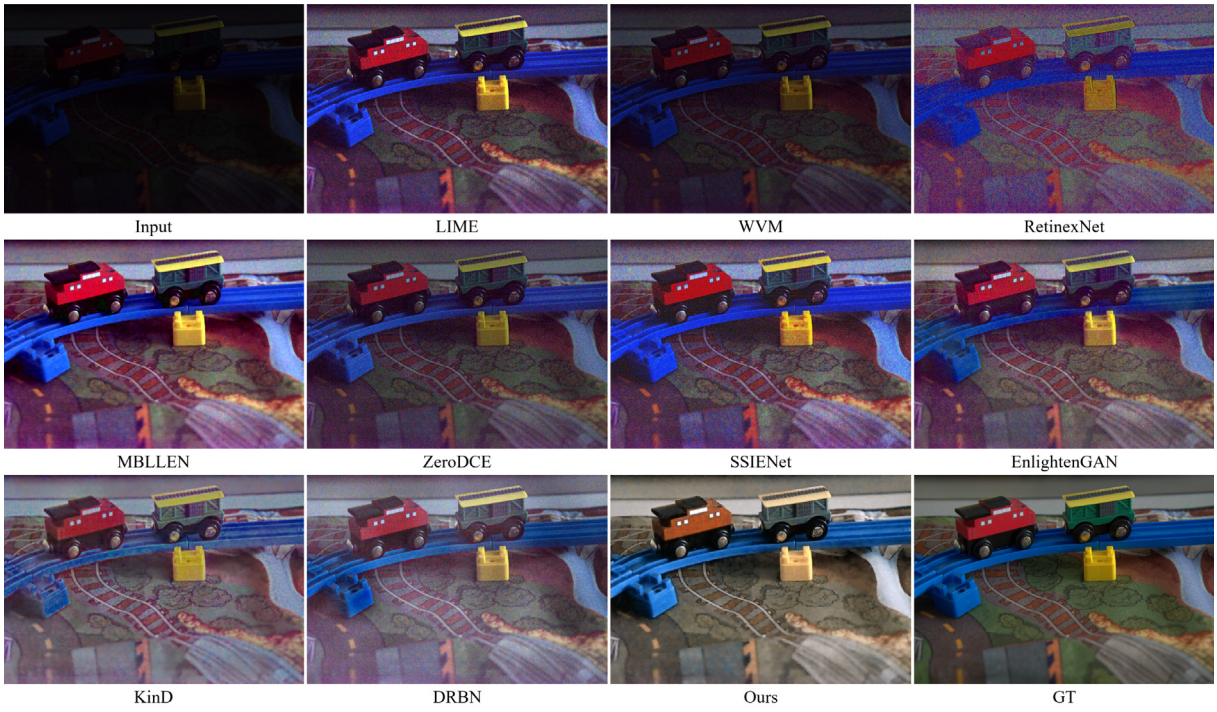
color bias and noise. Our method achieves the same good visual results as KinD and DRBN without serious noise and color bias, which demonstrates the superiority of our method for qualitative comparison.

Fig. 4 shows the visual results of different methods on the MIT-Adobe FiveK dataset. Fig. 5 presents the visual comparison results on the SID datasets. Intuitively, the results demonstrate that our method can enhance low-light images and produce pleasing results with proper brightness, color consistency, and clear details.

4.2.3. Efficiency evaluation

In addition to comparing the performance of the methods, we also compared the computational complexity of the different models, including the number of model parameters, the number of

(a)



(b)

Fig. 5. Qualitative visual comparison results of the different methods on the SID dataset.

floating-point operations and the runtime. The model parameters (in M) are the number of trainable parameters and measure the spatial complexity of the model. The FLOPs (in G) represent the total number of floating point operations performed by the model and measure the time complexity of the model. Here, the FLOPs of the model are measured on images of size $3 \times 1200 \times 900$. The runtime (in second) measures the inference time of the model, and the time statistics are performed on the LOL dataset. The runtime is measured on a PC with a GeForce GTX TITAN V GPU and Intel i7-6800 K CPU. The results of the computational complexity comparisons are reported in Table 4, and some of the computational complexity data are obtained from [50]. As can be seen from the Table 4, our method has a smaller number of parameters, FLOPs and runtime compared to most other methods. As such, our method achieves better performance with fewer parameters and fast inference time, balancing enhanced performance with computational complexity.

4.3. Ablation study

In this subsection, we conduct ablation studies to verify the effectiveness of each component of our method. We focused on the role of the gradient branch and knowledge distillation. In addition, we validate the complementarity of bilateral grids and knowledge distillation.

We first set up *base* as our baseline network, which contains only the enhancement branch and is trained using reference loss, without gradient branch and knowledge distillation. We add different components to the baseline network, which are: (1) base + gard (canny): Canny gradient is added to the baseline network. (2) base + gard(sobel): add Sobel gradient to the baseline network. (3) base + gard(origin): Feed the origin image to the gradient branch. (4) base + grad(sobel)+kd: Combine Sobel gradient and knowledge distillation together for training. (5) teacher: teacher network with a larger network model.

Study of Gradient Branch. First, we verify the effectiveness of the gradient branch by comparing base and each base + grad. As shown in Table 5, adding the gradient branch to base always improves the PSNR and SSIM metrics to some extent, which indicates the effectiveness of the gradient branch. Next, we explore the effectiveness of various gradient operators, including Canny, Sobel, as well as using the original image as the gradient input. Canny is not suitable due to its time-consuming computation and the loss of much information in the binary results. Compared to the gradient, the original image input may make the information duplicated. Sobel uses Gaussian smoothing to suppress the noise of the gradient map and provide better results, so we choose the Sobel operator to get the gradient map.

Study of Distillation Learning. We then verify the validity of knowledge distillation by comparing base + grad(sobel) and base + grad(sobel)+kd. In Table 5, the SSIM value is improved by adding knowledge distillation from the teacher network. The participation of the teacher network is not required in the testing phase and does not increase the computational burden of the inference process. This demonstrates the effectiveness of knowledge distillation in low-light image enhancement, which can improve performance without modifying the network structure and increasing the computational burden.

Study of complementarity with bilateral grids. When dealing with high-resolution images, the computational complexity increases squarely with the spatial resolution of the input image. HDRNet [51] uses a bilateral network technique with high computational efficiency. HDRNet first predicts the affine transform coefficients of the bilateral grid in low-resolution space, upsamples the coefficients to the original resolution size according to the guidance map, and finally applies the affine transform to the original image to enhance it. The main factors that affect the performance of the bilateral grid method include three aspects, the size of the coefficient prediction network, the spatial resolution of the grid and the depth of the grid, and changing these three factors increases the performance of the algorithm while also increasing the overhead of the algorithm.

Here, we focus on exploring the complementary properties of bilateral grids and distillation learning. The experiments are conducted on the MIT dataset with the network structure consistent with HDRNet, including the local branch, the global branch and the fusion module, and the guidance map generation uses a pixel-level shallow network and the loss function is L1 loss. The bilateral grid size of the student network is set to 4 and the spatial resolution is set to 8×8, while the corresponding parameters of the teacher network are set to 8 and 16×16, with the number of feature channels of the coefficient prediction network adjusted to be twice that of the student. As shown in the experiments in Table 6, after distilling the grid coefficients of the teacher network

Table 4
Quantitative comparison of the computational complexity of different Method. The evaluation metrics include Parameters (in M), FLOPs (in G), Runtime (in second).

	Parameters↓	FLOPs↓	Runtime↓
LIME	-	-	0.0710
WVM	-	-	4.9176
MBLLEN	0.450	301.120	1.4126
RetinexNet	0.555	587.470	0.5537
SSIENet	0.682	-	0.4280
ZeroDCE	0.079	84.990	0.0012
EnlightenGAN	8.637	273.240	0.0190
KinD	8.160	574.954	0.5388
DRBN	0.577	196.359	0.0132
Teacher	2.151	605.828	0.0228
Ours	0.283	83.953	0.0080

Table 5
Quantitative results of ablation studies.

	PSNR↑	SSIM↑
base	21.06	0.8068
base + grad(canny)	21.10	0.8072
base + grad(origin)	21.15	0.8096
base + grad(sobel)	21.19	0.8117
base + grad(sobel)+kd	21.24	0.8137
teacher	22.72	0.8350

Table 6
Quantitative results of complementary studies with bilateral grids.

	PSNR↑	SSIM↑
HDRNet	24.07	0.8862
HDRNet + kd	24.37	0.8892
HDRNetTeacher	24.84	0.8930

to the student HDRNet, its PSNR and SSIM are improved, which fully demonstrates that distillation learning can also improve the performance of the bilateral grid, which is complementary to the bilateral grid in model compression.

5. Conclusion

In this work, we propose a knowledge distillation method for low light image enhancement. The teacher network learns rich knowledge of the enhanced images, while the student network is simultaneously guided by the ground truth images and the teacher network. In addition, to better retain details, we design a gradient-guided network structure where enhancement branches are learned under the guidance of gradient branches. Our approach achieves a balance between computational complexity and performance, and improves the performance of the student network without increasing the computational burden during the testing phase. The experimental results demonstrate the superiority of our method over other state-of-the-art methods.

CRediT authorship contribution statement

Ziwen Li: Conceptualization, Methodology, Software, Writing – original draft. **Yuehuan Wang:** Supervision, Project administration, Funding acquisition. **Jinpu Zhang:** Resources, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

aaa

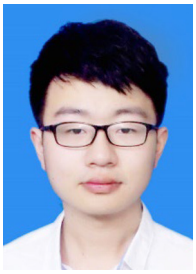
References

[1] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, 2014 IEEE Conference on Computer Vision and Pattern Recognition (2014) 1701–1708, https://doi.org/10.1109/CVPR.2014.220.
[2] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 815–823, https://doi.org/10.1109/CVPR.2015.7298682.
[3] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495, https://doi.org/10.1109/TPAMI.2016.2644615.
[4] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, IEEE Trans. Pattern Anal. Mach. Intell. 42 (2) (2020) 386–397, https://doi.org/10.1109/TPAMI.2018.2844175.
[5] K. Zuiderveld, Contrast limited adaptive histogram equalization, Graphics gems (1994) 474–485.
[6] J. Stark, Adaptive image contrast enhancement using generalizations of histogram equalization, IEEE Trans. Image Process. 9 (5) (2000) 889–896, https://doi.org/10.1109/83.841534.
[7] C. Lee, C. Lee, C.-S. Kim, Contrast enhancement based on layered difference representation of 2d histograms, IEEE Trans. Image Process. 22 (12) (2013) 5372–5384, https://doi.org/10.1109/TIP.2013.2284059.
[8] D. Coltuc, P. Bolon, J.-M. Chassery, Exact histogram specification, IEEE Trans. Image Process. 15 (5) (2006) 1143–1152, https://doi.org/10.1109/TIP.2005.864170.
[9] J.-T. Lee, C. Lee, J.-Y. Sim, C.-S. Kim, Depth-guided adaptive contrast enhancement using 2d histograms, 2014 IEEE International Conference on Image Processing (ICIP) (2014) 4527–4531, https://doi.org/10.1109/ICIP.2014.7025918.
[10] E.H. Land, The retinex theory of color vision, Scientific Am. 237 (6) (1977) 108–129.
[11] D. Jobson, Z. Rahman, G. Woodell, Properties and performance of a center/surround retinex, IEEE Trans. Image Process. 6 (3) (1997) 451–462, https://doi.org/10.1109/83.557356.
[12] D. Jobson, Z. Rahman, G. Woodell, A multiscale retinex for bridging the gap between color images and the human observation of scenes, IEEE Trans. Image Process. 6 (7) (1997) 965–976, https://doi.org/10.1109/83.597272.
[13] X. Guo, Y. Li, H. Ling, Lime: Low-light image enhancement via illumination map estimation, IEEE Trans. Image Process. 26 (2) (2017) 982–993, https://doi.org/10.1109/TIP.2016.2639450.
[14] X. Guo, Lime: A method for low-light image enhancement, in: Proceedings of the 24th ACM International Conference on Multimedia, MM '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 87–91. doi:10.1145/2964284.2967188.
[15] S. Wang, J. Zheng, H.-M. Hu, B. Li, Naturalness preserved enhancement algorithm for non-uniform illumination images, IEEE Trans. Image Process. 22 (9) (2013) 3538–3548, https://doi.org/10.1109/TIP.2013.2261309.
[16] M. Li, J. Liu, W. Yang, X. Sun, Z. Guo, Structure-revealing low-light image enhancement via robust retinex model, IEEE Trans. Image Process. 27 (6) (2018) 2828–2841, https://doi.org/10.1109/TIP.2018.2810539.
[17] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, J. Paisley, A fusion-based enhancing method for weakly illuminated images, Signal Process. 129 (2016) 82–96, https://doi.org/10.1016/j.sigpro.2016.05.031.
[18] X. Fu, D. Zeng, Y. Huang, X.-P. Zhang, X. Ding, A weighted variational model for simultaneous reflectance and illumination estimation, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2782–2790, https://doi.org/10.1109/CVPR.2016.304.
[19] K.G. Lore, A. Akintayo, S. Sarkar, Llnet: A deep autoencoder approach to natural low-light image enhancement, Pattern Recogn. 61 (2017) 650–662, https://doi.org/10.1016/j.patcog.2016.06.008.
[20] C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement, in: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3–6, 2018, 2018.
[21] F. Lv, F. Lu, J. Wu, C. Lim, MBLLEN: low-light image/video enhancement using CNNs, in: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3–6, 2018, 2018, p. 220.
[22] Y. Zhang, J. Zhang, X. Guo, Kindling the darkness: A practical low-light image enhancer, in: Proceedings of the 27th ACM International Conference on Multimedia, MM '19, Association for Computing Machinery, 2019, p. 1632–1640. doi:10.1145/3343031.3350926.
[23] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, Z. Wang, EnlightenGAN: Deep light enhancement without paired supervision, IEEE Trans. Image Process. 30 (2021) 2340–2349, https://doi.org/10.1109/TIP.2021.3051462.
[24] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531.
[25] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, K. Ma, Be your own teacher: Improve the performance of convolutional neural networks via self distillation, 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 3712–3721, https://doi.org/10.1109/ICCV.2019.00381.
[26] S. Zagoruyko, N. Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, arXiv preprint arXiv:1612.03928.
[27] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, J. Wang, Structured knowledge distillation for semantic segmentation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2599–2608, https://doi.org/10.1109/CVPR.2019.00271.
[28] Y. Liu, C. Shu, J. Wang, C. Shen, Structured knowledge distillation for dense prediction, IEEE Trans. Pattern Anal. Mach. Intell. (2020) 1, https://doi.org/10.1109/TPAMI.2020.3001940.
[29] W. Wang, C. Wei, W. Yang, J. Liu, Gladnet: Low-light enhancement network with global awareness, in: 2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018), 2018, pp. 751–755. doi:10.1109/FG.2018.00118.
[30] R. Wang, Q. Zhang, C.-W. Fu, X. Shen, W.-S. Zheng, J. Jia, Underexposed photo enhancement using deep illumination estimation, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 6842–6850, https://doi.org/10.1109/CVPR.2019.00701.
[31] L.-W. Wang, Z.-S. Liu, W.-C. Siu, D.P.K. Lun, Lightening network for low-light image enhancement, IEEE Trans. Image Process. 29 (2020) 7984–7996, https://doi.org/10.1109/TIP.2020.3008396.
[32] F. Zhang, Y. Li, S. You, Y. Fu, Learning temporal consistency for low light video enhancement from single images, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4965–4974, https://doi.org/10.1109/CVPR46437.2021.00493.
[33] R. Liu, L. Ma, J. Zhang, X. Fan, Z. Luo, Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10561–10570.
[34] Q. Yang, Y. Wu, D. Cao, M. Luo, T. Wei, A lowlight image enhancement method learning from both paired and unpaired data by adversarial training, Neurocomputing 433 (2021) 83–95.
[35] Z. Jiang, H. Li, L. Liu, A. Men, H. Wang, A switched view of retinex: Deep self-regularized low-light image enhancement, Neurocomputing 454 (2021) 361–372.
[36] C. Guo, C. Li, J. Guo, C.C. Loy, J. Hou, S. Kwong, R. Cong, Zero-reference deep curve estimation for low-light image enhancement, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 1777–1786, https://doi.org/10.1109/CVPR42600.2020.00185.

[37] W. Yang, S. Wang, Y. Fang, Y. Wang, J. Liu, From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3060–3069, https://doi.org/10.1109/CVPR42600.2020.00313.
[38] T. Chen, I. Goodfellow, J. Shlens, Net2net: Accelerating learning via knowledge transfer, arXiv preprint arXiv:1511.05641.
[39] S. Lin, R. Ji, C. Chen, D. Tao, J. Luo, Holistic CNN compression via low-rank decomposition with knowledge transfer, IEEE Trans. Pattern Anal. Mach. Intell. 41 (12) (2019) 2889–2905, https://doi.org/10.1109/TPAMI.2018.2873305.
[40] Q. Li, S. Jin, J. Yan, Mimicking very efficient network for object detection, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 7341–7349, https://doi.org/10.1109/CVPR.2017.776.
[41] L. Zhang, K. Ma, Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, OpenReview.net, 2021.
[42] Q. Gao, Y. Zhao, G. Li, T. Tong, Image super-resolution using knowledge distillation, in: Computer Vision – ACCV 2018, Springer International Publishing, Cham, 2019, pp. 527–541.
[43] Z. He, T. Dai, J. Lu, Y. Jiang, S.-T. Xia, Fakd: Feature-affinity based knowledge distillation for efficient image super-resolution, 2020 IEEE International Conference on Image Processing (ICIP) (2020) 518–522, https://doi.org/10.1109/ICIP40778.2020.9190917.
[44] W. Lee, J. Lee, D. Kim, B. Ham, Learning with privileged information for efficient image super-resolution, 16th European Conference Computer Vision ECCV, vol. 12369, Springer, 2020, pp. 465–482, https://doi.org/10.1007/978-3-030-58586-0_28.
[45] M. Hong, Y. Xie, C. Li, Y. Qu, Distilling image dehazing with heterogeneous task imitation, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3459–3468, https://doi.org/10.1109/CVPR42600.2020.00352.
[46] X. Qin, Z. Wang, Y. Bai, X. Xie, H. Jia, FFA-Net: Feature fusion attention network for single image dehazing, in: Proceedings of the AAAI Conference on Artificial Intelligence(AAAI), 2020, pp. 11908–11915.
[47] V. Bychkovsky, S. Paris, E. Chan, F. Durand, Learning photographic global tonal adjustment with a database of input/ output image pairs, in: CVPR 2011, 2011, pp. 97–104. doi:10.1109/CVPR.2011.5995332.
[48] C. Chen, Q. Chen, J. Xu, V. Koltun, Learning to see in the dark, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018) 3291–3300, https://doi.org/10.1109/CVPR.2018.00347.
[49] Y. Zhang, X. Di, B. Zhang, C. Wang, Self-supervised image enhancement network: Training with low light images only, arXiv preprint arXiv:2002.11300.
[50] C. Li, C. Guo, L.-H. Han, J. Jiang, M.-M. Cheng, J. Gu, C.C. Loy, Low-light image and video enhancement using deep learning: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1, https://doi.org/10.1109/TPAMI.2021.3126387.
[51] M. Gharbi, J. Chen, J.T. Barron, S.W. Hasinoff, F. Durand, Deep bilateral learning for real-time image enhancement, ACM Trans. Graph. 36 (4). doi:10.1145/3072959.3073592.

Ziwen Li is currently a Ph.D. candidate in the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, supervised by Prof. Yuehuan Wang. His research interests include image enhancement and deep learning.

Yuehuan Wang graduated from University of Electronic Science and Technology of China in 1993, and received an M. S. degree in computer system architecture and a Ph.D. degree in pattern recognition and Artificial intelligence from Huazhong University of Science and Technology, China, in 1996 and 2001, respectively. He has been a visiting scholar working with professor Jun Shen in University of Bordeaux III, and with professor Amir A Amini in Washington University in St Louis. He is currently a professor with School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. His research interests include computer vision, image understanding, target tracking, and automatic target recognition.



Jinpu Zhang is currently pursuing the Ph.D. degree in pattern recognition and intelligent systems from Huazhong University of Science and Technology. His research interests include visual tracking, object detection and computer vision.