Low Lumination Image Enhancement with Transformer based Curve Learning

Yulin Cao[1], Chunyu Li[2], Guoqing Zhang[2], and Yuhui Zheng[1*]
[1] Colloge of computer, Qinghai Normal University
Xining, Qinghai, China
[e-mail: zhengyh@vip.126.com]
[2] School of Computer and Software, Nanjing University of Information Science and Technology
Nanjing, China
[e-mail: 72421@163.com]
[*]Corresponding author: Yuhui Zheng

Received December 8, 2023; revised July 4, 2024; accepted August 8, 2024; published September 30, 2024

Abstract

Images taken in low lamination condition suffer from low contrast and loss of information. Low lumination image enhancement algorithms are required to improve the quality and broaden the applications of such images. In this study, we proposed a new Low lumination image enhancement architecture consisting of a transformer-based curve learning and an encoder-decoder-based texture enhancer. Considering the high effectiveness of curve matching, we constructed a transformer-based network to estimate the learnable curve for pixel mapping. Curve estimation requires global relationships that can be extracted through the transformer framework. To further improve the texture detail, we introduced an encoder-decoder network to extract local features and suppress the noise. Experiments on LOL and SID datasets showed that the proposed method not only has competitive performance compared to state-of-the-art techniques but also has great efficiency.

Keywords: Enhancement, transformer, Low lumination, spatial attention, channel attention, convolutional neural network.

1. Introduction

The digital image has become an important carrier of information and plays an essential role in our daily lives. However, the quality of images usually suffers from insufficient or uneven illumination, resulting in noise and low contrast images. This seriously hinders the application of digital images, such as mobile photography, video surveillance, and autopilot. Therefore, digital imaging should work in different scenarios including low lumination conditions. In order to make better use of the information contained in images, there is an urgent need to develop an efficient low lumination image enhancement method.

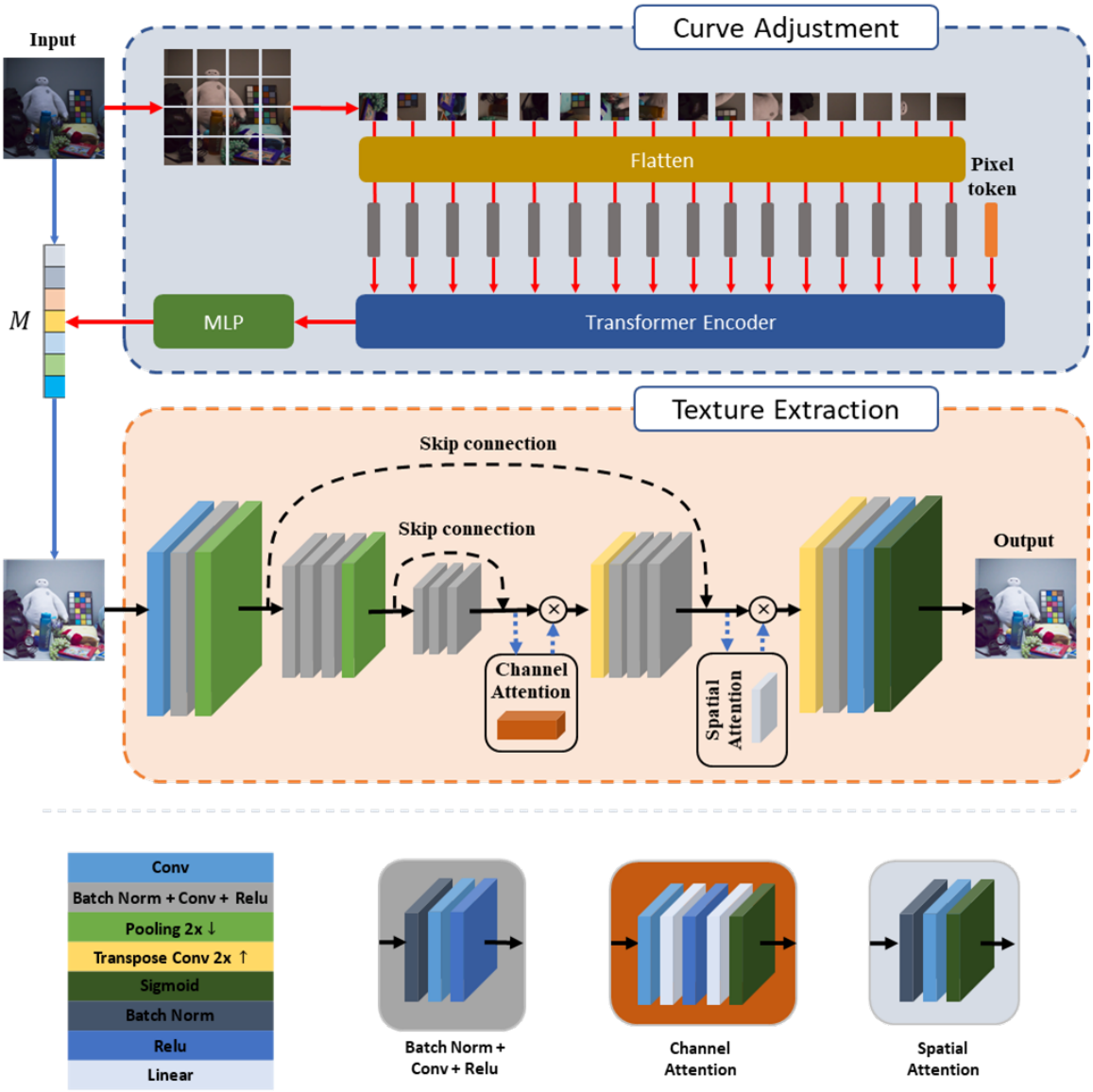

Fig. 1. The main framework of the proposed method, which consists of a curve adjustment module and a texture extraction module.

Over the past decades, a variety of enhancement methods have been introduced. They can be roughly divided into four categories: histogram-based methods [1-5], retinex-based approaches [6-15], inverted domain-based techniques [16, 17], and data-driven methods [18-28]. The histogram equalization (HE)-based methods mainly focus on improving the contrast by adjusting the statistical distribution of individual pixels. They extend the dynamic range to recover detail [4]. Histogram-based approaches are usually time-efficient but tend to over-enhance the noise in homogeneous regions [1]. Retinex theory states that the image captured by the human eyes can be considered as the product of illumination and reflectance. The main processes of retinex-based methods include the decomposition of the illumination and the reflectance, the adjustment of the decomposed components by specific procedures, and the fusion of the adjusted components. The earliest techniques [6, 7] estimated the reflectance of the image and considered the reflectance as the enhanced result. They did not take the illumination into account, resulting in the probability of over-enhancement. Later research estimated the illumination of Low lumination images and the adjusted them. The enhanced result was obtained by multiplying the reflectance and the adjusted illumination. The inverted domain-based methods are inspired by the observation that the inverted Low lumination images share many characteristics with the image taken in hazy conditions. They apply a de-haze algorithm to the inverted image and then invert the de-hazed image to obtain the improved result. The frameworks for these techniques are manually designed and tend to suffer from a lack of feature representation, leading to unsatisfactory performance.

In recent years, deep learning-based approaches have been rapidly developed due to their superiority in feature extraction and representation. Specifically, they construct end-to-end frameworks to learn the signal features from Low lumination images and map them to the target representation in RGB or raw domain [18, 19]. The references of [20, 22, 25, 26, 28] combine retinex theory and deep learning, decompose the illumination and reflectance through convolutional neural networks, and finally tune them to obtain the improved results. In the reference [21], content and structure were learned through two connected streams, and a spatial variant was proposed to extract structure features. In the reference [23], a generative adversarial network framework was introduced to enhance the Low lumination images in the generative model. Li *et al.* [24] designed a high-order curve to map a Low lumination image to a normal-light image. The parameters of the curve were learned using a deep curve estimation network. The frameworks of the above methods are almost based on the convolutional neural network, which has advantages in modelling the local relationship of images but fails to extract the global features.

In this study, we proposed a novel framework including two subnetworks: a transformer-based curve adjustment network and a context enhancement and denoising network (**Fig. 1**). Since the contrast enhancement is a global operator but the context enhancement and denoising are mostly based on the local features, we solved them in independent sub-networks. The structure of the curve adjustment network was based on the vision transformer (VIT) [29] because of its excellent ability to establish a global relationship across the entire image. The context enhancement and noise reduction network was based on an encoder-decoder convolutional neural network which was able to extract local features and suppress noise. Furthermore, channel attention and spatial attention modules were introduced to improve performance.

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

1) We proposed a novel curve learning network to stretch the contrast and recover the information buried in the darkness, which was based on the transformer and was generalized well to different images.

2) We constructed an encoder-decoder-based texture enhancement network and integrated the attention mechanism, which explores the local features of the image and avoids amplifying the noise.
3) We introduced a new strategy for generating the reference curve for training and constructing a loss function with a second-order gradient regularization to improve the curve adjustment results

The remainder of this article is organized in the following sections: Section 2 gives a brief overview of previous popular investigations on Low lumination image enhancement. In section 3, we propose our architecture and give a detailed introduction. Experimental results and discussion are presented in Section 4. Finally, conclusions and future work are discussed in Section 5

2. Related Works

2.1 Hand-Crafted Methods

Histogram Equalization: Due to its simplicity and efficiency, HE is a widely used technology for improving the visibility of Low lumination images. HE-based methods have a solid foundation of statistical mathematics. They adjust each pixel to balance the probability distribution, thereby increasing the dynamic range and improving contrast. To overcome the over-enhancement of the bright region, which weakens the local detail, Adaptive Histogram Equalization (AHE) [1] adjusts the histogram in the local block of the image. And an interpolation between local blocks is used to maintain the continuity of the overall image. Although AHE can greatly improve contrast and structure, it cannot suppress the noise which severely destroys much of the information contained in the image. Noise can be reduced by Contrast Limited Adaptive Histogram Equalization (CLAHE) [2]. CLAHE applies a limit to the enhancement in homogeneous areas and truncates the histogram to avoid the amplification of noise. The brightness of the image is usually changed after HE processing, which limits its use in areas where the original brightness should be preserved such as TV. To overcome this problem, Brightness Preserving Bi-Histogram Equalization (BBHE) [3] has been proposed as an extension of HE. BBHE decomposes the image into two sub-images based on their mean and applies independent HE to them. This imposes a constraint between the sub-images, resulting in the brightness level being bounded around the mean of the input.

Later, methods were proposed to improve HE performance using contextual information. For instance, the Contextual and Variational Contrast Enhancement (CVC) technique [5] constructed a 2-D histogram based on the relationship between pixels in the local region and imposed a constraint between the input histogram and the uniformly distributed histogram. In addition, the Layered Difference Representation (LDR) approach [5] improved the output differences using the LDR of a 2-D histogram which was based on the statistical information between neighboring pixels.

However, HE-based methods tend to focus on improving the contrast of the overall image, rather than exploiting and enhancing the illumination of the Low lumination image, and therefore under the risk of over- or under-enhancement.

Retinex-based Methods: The Retinex theory [30] suggests that an image that we observe can be decomposed into two components: illumination and reflectance, and that the colour of an object is determined by reflectance. Based on the Retinex theory, a series of approaches have been proposed. Single-Scale Retinex (SSR) [6] assumes that the illumination is smooth, and estimates it using a Gaussian filter. Then The illumination is then removed and the

reflectance is used as the enhanced result. It is difficult to maintain a balance between colour and dynamic range using SSR, and therefore, so Multi-scale Retinex (MSR) [7] was proposed to alleviate this problem. MSR applies a Gaussian filter with different scales and calculates the average of these filtered outputs. Compared to SSR, MSR can simultaneously achieve colour reproduction and dynamic range compression. However, SSR and MSR remove the illumination directly from the original image and treat the reflectance as the enhancement output, which destroys the naturalness and tends to over-enhance the result.

Recently, the retinex-based methods mainly decompose the illumination and reflectance and enhance them using various techniques, and then treat the fusion version of them as the enhanced result. For instance, MF [13] performs decomposition using morphological closure, and the enhancement is performed based on sigmoid function and AHE. The adjusted illumination is constructed by a multi-scale fusion, and then it is multiplied by the estimated reflectance to obtain the final result. Guo *et al.* [8] estimated the illumination in a variational model, which found the maximum pixel value in RGB channels to construct the initial illumination map as the reference and introduced a structure-aware weight to smooth the illumination. To improve the structure and texture of the enhancement result, STAR [10] used exponentiated local derivatives to extract the structure and texture maps and used them to control the illumination and reflectance. In the references of [9, 12] and [15], the noise was explicitly considered and a noise term was added to the original retinex theory. This helps to improve the performance of the enhancement for the Low lumination image, which always contains noise that is simultaneously amplified during the brightness enhancement.

Although retinex-based approaches perform well in terms of illumination enhancement and detail detection, they fail to preserve the naturalness which can affect the visual perception.

Inverted domain-based techniques: Dong *et al.* [17] observed that the inverted version of a Low lumination image has many similar characteristics compared to an image captured under hazy conditions. They inverted the Low lumination image and then applied a dehazing algorithm to the inverted image, called video dehazing. The improved result was obtained by inverting the dehazed image again. In the reference [16], an improved dehazing algorithm was proposed to dehaze the inverted image by estimating the transmission based on the luminance instead of depth. Inverted domain-based methods are fast and efficient, but they do not produce satisfactory visual quality. Furthermore, there is no reasonable explanation for their effectiveness.

2.2 Data-Driven Methods

With the rapid development of deep learning, various data-driven methods have been proposed for low lumination image enhancement. The encoder-decoder architecture is a widely adopted framework in various domains. For example, Lore *et al.* [18] constructed an autoencoder (LLNet) that was trained to extract signal features from the Low lumination image and to improve the contrast. LLNet focuses on the local patch-wise contrast enhancement based on neighbors, and suppresses noise by another denoising autoencoder. Ren *et al.* [21] approached the contrast enhancement through two streams, focusing on context and edge, respectively. The context stream consists of an encoder-decoder network, which is designed to estimate the global content in the Low lumination image. The edge stream uses the same encoder-decoder network but predicts the edge using spatially varying recurrent neural networks (RNNs). Unlike the previously mentioned methods, the reference [19] presented an end-to-end convolutional neural network (CNN) that directly performs an enhancement using the raw sensor data and, discarding the traditional pipeline.

Shen *et al.* [31] found that the MSR is equivalent to an end-to-end CNN consisting of multiple Gaussian convolutional kernels. And a convolutional neural network (MSR-net) with a similar architecture to the MSR was proposed to map a Low lumination image to a normal-light image. Then, more and more methods combining CNN and Retinex were proposed. For example, Li *et al.* [28] proposed trainable CNN (LightenNet) to estimate the illumination and produce the enhanced image by referring to the Retinex model. Wei et al. [20] constructed a three-step framework. In this framework, illumination and reflectance were decomposed using a Decom-Net, then enhanced using an encoder-decoder network. Finally, the enhanced result was reconstructed based on the retinex model. Deng *et al.* [22] built a joint decomposition and denoising network based on the U-Net architecture[32]. The network was trained under the assumption that the reflectance map is similar to the paired normal-light image, which would simultaneously suppress the noise in the reflectance. The RetinexDIP [26] describes the retinex decomposition as a generative problem. With reference to the Deep Image Prior (DIP) [33], RetinexDIP generates the reflectance and illumination from stochastic noise using two dividual DIP networks.

Generative adversarial networks (GANs) perform well in image enhancement and are free of paired data. Kim *et al.* [27] first proposed a GANs-based algorithm for Low lumination image enhancement, which was trained with adversarial loss, perceptual loss, and colour loss. To avoid over- or under-enhancement in local regions, EnlightenGAN [23] constructed the generator network with an attention mechanism that used illumination to construct an attention map. Global-local discriminators were also considered to improve the performance with no-uniform illumination.

3. Proposed Method

We proposed a novel architecture based on transformer curve learning and CNN to perform Low lumination image enhancement. The framework is divided into two steps: brightness adjustment based on transformer curve learning and context enhancement based on encoder-decoder CNN. The transformer can establish the non-local relationship that contributes to the curve adjustment on the whole image. And the context enhancement module is introduced to extract the local detail and suppress the noise. **Fig. 1** illustrates the structure of our proposed methods.

3.1 Curve Adjustment with Transformer

The curve adjustment in image processing software shows great performance in Low lumination image enhancement with hand-craft settings. The curve represents the mapping of pixel values between the Low lumination and normal-light images. However, it is limited in practical applications because the settings are only appropriate for the specific image. To overcome this problem, we tried to learn a curve using deep neural networks.

Compared to the gamma correction, curve adjustment is a more universal algorithm for improving image brightness because the shape of the curve is more flexible and the mapping can fit images with different dynamic ranges (**Fig. 2(a)**). The curve is a function that maps one pixel to another. Taking the 8-bit image as an example, the domain of the function can be represented as $\{x|\ 0 \leq x \leq 255\}$ and the range as $\{y|\ 0 \leq y \leq 255\}$. For the pixel values which were integer and independent, we formulated a discrete function to represent the mapping and to impose smooth constraints.

As we know, the curve transformation is a global operation, which is based on the non-local relationship between pixels. The CNN, which has been widely used in image

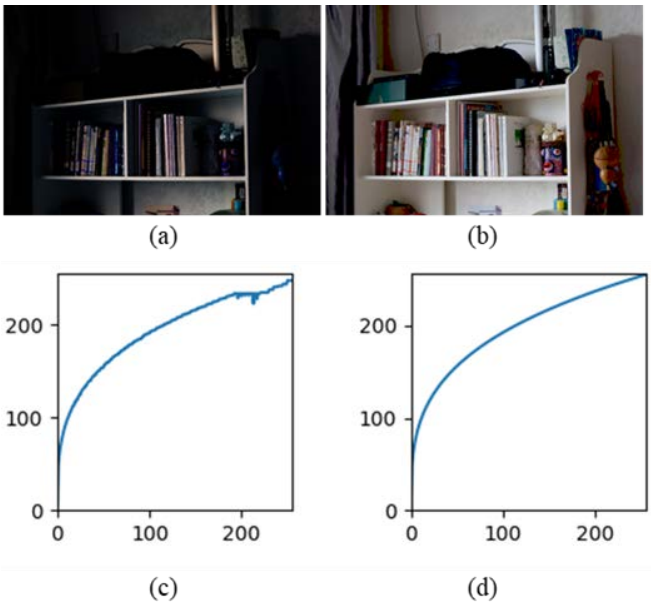

Fig. 2. An example of curve adjustment: (a) Low lumination image; (b) the output of curve adjustment; (c) the curve used to enhance the low-light image; (d) the calculated reference curve.

enhancement, focuses mainly on the extraction of local features but fails to establish the global relationship. Knowing that vision transformer (ViT) [29] has achieved great success in various computer vision tasks, which can extract the global features, we constructed a transform framework to estimate the mapping.

A transformer is a novel structure based on self-attention, which is originally proposed for natural language processing (NLP) [34]. Dosovitskiy *et al.* [29] first introduced a transformer to computer vision. In ViT, images are divided into patches and are flattened as vectors. Multi-head self-attention modules and multi-layer perceptron (MLP) are employed to extract features after adding the learning classification token.

Inspired by the ViT framework, we constructed the learnable pixel token to replace the classification token and designed a curve prediction head to generate the map of pixels, representing the corresponding values in the adjusted result. The map was represented using the following format:

$$M = [p_0,\ p_1,,\ldots,p_{256}] \tag{1}$$

The elements $p_0,\ p_1,,\ldots,p_{256}$ denote the corresponding values of the original values $[1,2,\ldots,256]$ in the adjusted result. The transformer-based curve learning framework is illustrated in **Fig. 1**. The Low lumination image is divided into 2D patches using a convolutional kernel of size 16 and then is flattened as one-dimensional tensors. After stacking with a pixel token, a transformer encoder is constructed to extract features from the input patches which contain several transformer blocks. The blocks consist of multi-head self-attention modules [29] and linear layers that are followed by activation and layer normalization.

To train the network that could be applied to different images, we constructed the ground truth using the paired low lumination and normal-light data. As a reference, we took the dominant value in all the pixels in the normal light image that corresponded to the specific pixel value in the low lumination image. **Fig. 2(d)** illustrates an example of reference calculated between **Fig. 2(a)** and **(b)**, where the shape is similar to the ground truth curve in **Fig. 2(c)**. This is a clear demonstration of the validity of this reference construction algorithm.

In the test stage, the Low lumination image is fed into our transformer-based curve estimation framework, and we will get the map of pixels. The improved results are obtained by discrete curve adjustment.

The advantage of using transformers lies in their self-attention mechanism, allowing the network to process long-range dependencies and establish global contexts on images, unlike traditional convolutional neural networks that are more adept at capturing local features.

3.2 Context Enhancement and Denoising

The global curve adjustment introduced above significantly improved the contrast of the Low lumination image and revealed the objects buried in the dark areas. However, at the same time, the noise was amplified simultaneously and the global curve adjustment was unable to capture the local detail. We addressed the challenges associated with enhancing low luminance images by employing an encoder-decoder network to enhance texture details and suppress noise, along with channel and spatial attention modules to further improve performance.

Due to the great performance of encoder-decoder networks in image denoising [35-37], deraining [38], dehazing [39], and inpainting [40], we constructed our context enhancement and denoising network based on the encoder-decoder architecture. The context extractor consists mainly of convolutional layers, batch norm layers, pooling layers, and transpose convolutional layers. The encoder acts as a feature actor that learns feature maps in local regions through several convolutional layers. Pooling layers are used to down-sample the features and increase the receptive field. The bottleneck block, consisting of three convolutional layers, performs a non-linear map on the extracted features. The decoder network fuses the features and reconstructs the output, and learnable transpose convolutional layers are introduced to up-sample the features.

Unlike high-level tasks such as classification, object tracking, and segmentation which rely on high-level features, image restoration requires low-level features which are typically generated from the shallow layers in networks. To take advantage of the low-level features, we introduced skip connections between the blocks of our encoder-decoder structure. **Fig. 1** exhibits that skip connections are added at the end of each block of the encoder so that features of different levels can be transferred to corresponding blocks of the decoder. In addition, the introduction of skip connections is helpful for network convergence.

With the skip connections, the low-level features are stacked with the corresponding high-level features, but some features may be redundant. In order to make the best of the stacked features, we introduced a channel attention module that can assign weights to different channels of the features. The channel attention module takes an adaptive average pooling layer as the first layer to enlarge the receptive field, then linear layers and activation layers are used to compute the weights. The channel attention mechanism causes our network to pay more attention to the features with more importance. Before the last block of our decoder, which produces the improved results, a spatial attention module is added. The spatial attention module consists of a convolutional layer and a sigmoid layer. It is a pixel-wise weighting approach that overcomes the limitation that the curve adjustment using a single curve cannot adequately enhance different regions of the image.

3.3 Loss Function

Since the Contexture extractor took the curve-adjusted image as the input, we had to train our network in two steps. First, the transformer-based curve learning network was trained first, then a brightness enhancement was performed on the original Low lumination input after which the contexture extractor could be trained. We defined loss functions for the two-

component networks.

Curve adjustment Loss: The output of our curve adjustment module is a vector of 256 elements. To measure the difference between the estimated curve and the reference curve, we used the mean absolute error (MAE), which is insensitive to outliers and the gradient was more stable. The MAE loss can be written as follows:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{n}\sum_{i=1}^{n}\left|M_i - M_i^{ref}\right|, \tag{2}$$

where M_i and M_i^{ref} represent the i-th elements of the predicted curve and reference curve.

To improve the naturalness of the enhanced image, we proposed an element-wise weighted Laplace regularization term to smooth the estimated curve. It is defined as follows:

$$\mathcal{L}_{\text{Smooth}} = W \cdot (M_{i-1} + M_{i+1} - 2M_i), \tag{3}$$

$$W = (M_i/255)^{\gamma}. \tag{4}$$

M_{i-1} and M_{i+1} indicate the previous and next element of the current element (M_i) in the curve vector, and W stands for the element-wise weights. The Laplacian operator is a second-order operator. When it is used as a regularization, it imposes a constraint on the rate of change of the slope of the estimated curve rather than the slope itself. This can synchronously smooth the curve and preserve its ability to stretch the contrast, especially when $\gamma < 1$.

By combining the MAE loss and smooth loss, the final curve adjustment loss is defined as:

$$\mathcal{L}_{\text{curve}} = \mathcal{L}_{\text{MAE}} + \lambda_s \cdot \mathcal{L}_{\text{smooth}}, \tag{5}$$

where λ_s is the balance weight between $\mathcal{L}_{\text{MAE}}$ and $\mathcal{L}_{\text{smooth}}$.

Context Enhancing and Denoising Loss: The Contexture extractor was proposed to improve the context and to suppress the noise. We minimized the MAE loss to train the network to extract details from the input image. Furthermore, a total variant (TV) regularization was imposed on the output for denoising. The loss function can be written as:

$$\mathcal{L}_{\text{CED}} = \left\|I - I^{ref}\right\|_1 + \lambda_t \cdot TV(I), \tag{6}$$

I and I^{ref} represent the estimated result and ground truth image of our Contexture extractor. $TV(\cdot)$ is a total variant regularization term and λ_t is the balance weight.

4. Experiment

In this stage of our study, we set up experiments and compared the proposed method with state-of-the-art techniques from different aspects. Furthermore, the parameters and effects of different parts of the proposed framework were analyzed.

4.1 Experiment Setup

The experiments were implemented using PyTorch and were conducted on a PC with an Intel Core i9-1080XE CPU, an RTX 3090 GPU, and 192 GB of memory. The trade-off parameters of λ_s and λ_t in loss functions were set to 0.001 and 0.0005, respectively. The maximum iterations of the transformer-based curve adjustment module and texture extraction module were set to 20 and 50, respectively. To evaluate the performance of the proposed method, we conducted experiments on datasets including LOL [20] and SID [19]. LOL dataset provides 500 paired low/normal-light images taken from real scenes for Low lumination enhancement [20]. SID is another dataset that provides image pairs, but its raw images require preprocessing to obtain the RGB images.

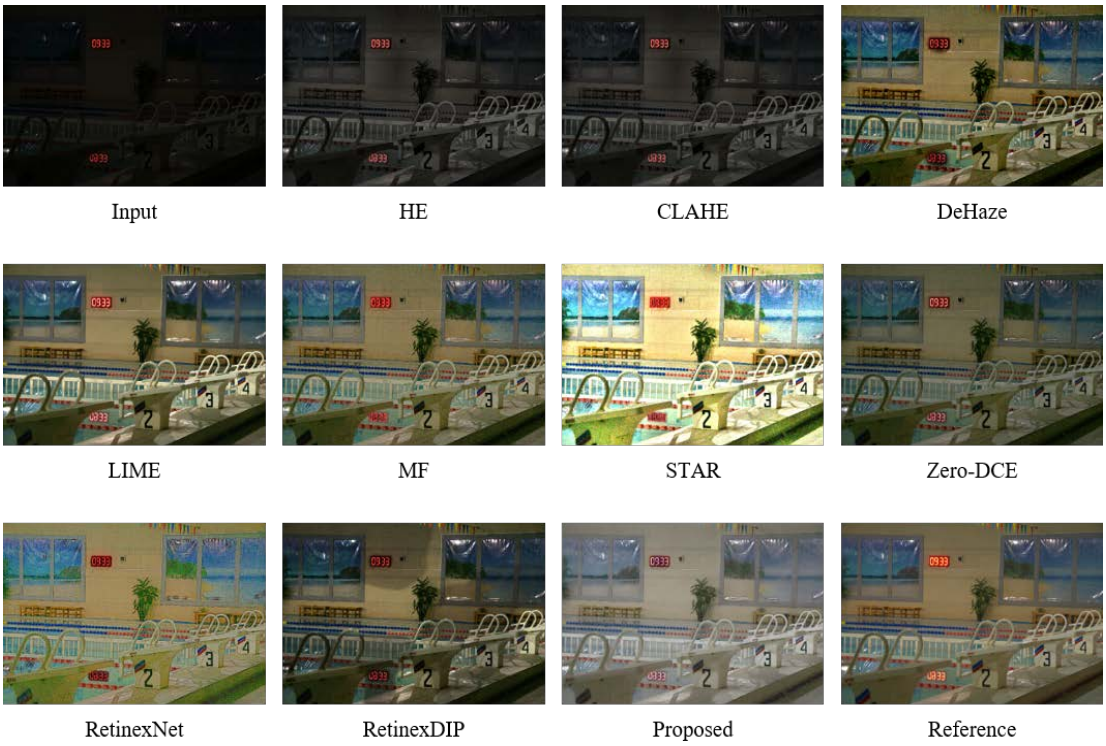

Fig. 3. The visual comparison of results produced by different methods using an image from LOL dataset.

Table 1. Quantitative Results of different Methods On LOL Dataset

	HE	CLAHE	DeHaze	LIME	MF	STAR	Zero-DCE	RetinexNet	RetinexDIP	Proposed
PSNR	13.4129	12.3891	16.978	16.9115	16.9075	18.8411	16.8718	8.8826	16.507	**25.1311**
SSIM	0.47884	0.4219	0.56641	0.69663	0.69666	0.76828	0.73847	0.50373	0.73616	**0.82069**
NIQE	5.6898	4.0885	4.7458	4.8034	4.7958	3.8571	4.6816	3.738	4.4082	**3.3481**

Table 2. Quantitative Results of different Methods On SID Dataset

	HE	CLAHE	DeHaze	LIME	MF	STAR	Zero-DCE	RetinexNet	RetinexDIP	Proposed
PSNR	13.2097	10.0781	16.2768	17.1818	16.9662	11.206	14.8607	16.774	9.96	**19.8184**
SSIM	0.3249	0.3104	0.5295	0.6349	0.6049	0.4625	0.5849	0.5594	0.3263	**0.7526**
NIQE	10.1137	8.36	8.9128	3.9584	9.7125	8.233	8.2232	9.7296	6.8967	**3.0527**

Our framework consists of the curve adjustment module and the texture extraction module. In the first phase, the curve adjustment module was trained. Since we trained the curve adjustment module in supervised mode, the reference curve vector was required which was introduced in section 3.A. The curve adjustment module takes the Low lumination image as the input and flattens the image patches into 1D sequences. After being embedded in the learnable curve token, the sequences are input to the transformer encoder to extract features. The output curve can be obtained via the MLP decoder. After performing a curve adjustment

on the input low lumination image using the obtained curve, we obtain the result of the contrast-enhanced version. In the second stage, we used the texture extraction module to improve the context information and to denoise. It takes the contrast result as input and the paired normal-light image as ground truth. The encoder-decoder structure learns the feature maps and reconstructs the final enhanced result. The curve adjustment module and texture extraction module share the enhancement tasks and work together to produce a satisfactory enhanced result.

4.2 Comparison with other methods

We compared the performance of our method with representative or state-of-the-art methods from different categories including HE-based approaches: HE, CLAHE [2], and DeHaze [17]; Retinex-based methods: LIME [8], MF [13], and STAR [10]; Learning-based methods: RetinexNet [20], RetinexDIP [26], and Zero-DCE [41]. We used their official codes and the default parameters provided, and the tests were performed in the same environment.

Qualitative assessments: **Fig. 3** shows that visual comparisons were made with other methods on the LOL dataset. From **Fig. 3**, we can see that these methods successfully recover information buried in dark except for the HE-based methods. However, the results of Dehaze and RetinexDIP are still somewhat weak, while the results of STAR and RetinexNet are over-enhanced. There is a lot of noise in the results of STAR, RetinexNet, and MF. For the LIME, the result seems unreal due to the over-denoising. The result of the proposed method not only improves the brightness, but also detects the texture well without amplifying the noise.

Fig. 4 shows the comparison of an outdoor scene from the SID dataset. It indicates that STAR and RetinexNet failed to process the sky region and that the brightness is still low for RetinexDIP. The results of the presented model achieved a high agreement with the reference. The visual comparisons demonstrated the effectiveness of the proposed method.

Quantitative assessments: We took three popular assessment criteria to quantitatively compare the performance of the presented method with state-of-the-art approaches, including the peak signal-to-noise ratio (PSNR), the structural similarity (SSIM) [42], and the Natural Image Quality Evaluator (NIQE) [43]. PNSR is one of the most widely used quality metrics, which is based on the mean squared error. SSIM measures the structural similarity between the input and reference images, while NIQE is a non-reference metric that focus on the naturalness.

The quantitative comparison of the different methods on 15 test images from the LOL dataset is shown in **Table 1**. The best results are marked in bold. The result of PSNR indicates that the aforementioned methods are effective in information detection and enhancement. The retinex-based methods and the data-driven techniques achieved better performance than the HE-based approaches, and the proposed method achieved the best result. The basic idea of HE-based methods is to improve the contrast of images, but it fails to preserve the structure and details, resulting in poor performance in SSIM and NIQE. The retinex-based methods perform well in terms of SSIM, assuming that the illumination is smooth and the reflectance contains the structure information. The data-driven methods gave better performance in NIQE. By introducing our texture extraction module, which explores the local details and suppresses the noise, the proposed method had superiority in SSIM and NIQE.

The quantitative results of the SID dataset are shown in **Table 2**, demonstrating the high performance of the proposed method. The LOL dataset contains more outdoor scenes and while the SID dataset contains more indoor scenes. The results convincingly prove the universality of the proposed model.

Fig. 4. The visual comparison of results produced by different methods using an image from SID dataset.

Table 3. The Results of Ablation Experiments

Curve	Encoder-decoder	Attention	PSNR	SSIM	NIQE
✓	×	×	18.4327	0.6437	8.1877
✓	✓	×	19.4723	0.7397	3.2827
✓	✓	✓	**19.8184**	**0.7526**	**3.0527**

Table 4. Comparison of Computing Time of different Methods

	HE	CLAHE	DeHaze	LIME	MF	STAR	Zero-DCE	RetinexNet	RetinexDIP	Proposed
Time (S)	0.0215	0.0211	0.0394	0.3845	0.1225	3.7197	0.0475	0.2387	32.5141	0.3374

4.3 Ablation Experiments

To evaluate the effectiveness of each module, we conducted ablation experiments on the LOL dataset. Three settings were used for comparison including: *i*. curve adjusting with transformer only; *ii*. curve adjusting with encoder-decoder enhancer; *iii*. curve adjusting and encoder-decoder enhancer with attention module. The quantitative results of the average of 15 testing images are shown in **Table 3**. It can been seen that the curve adjusting itself (setting *i*) can produce good results, but the naturalness is poor. When the encoder-decoder enhancer is adopted (the setting *ii*), the three metrics improve significantly, especially the NIQE. This

demonstrates the importance of the encoder-decoder enhancer in texture improvement. **Table 3** shows that the whole architecture (setting *iii*) achieves a further improvement in these metrics, proving the effectiveness of introducing attention modules. The ablation experiments suggest that each part of the proposed model is essential for competitive performance.

4.4 Computational Complexity

The comparison of computation time with other methods is shown in **Table 4**. The resolution of the test images was 600×400 And the results showed that the HE-based methods were the most efficient methods. The retinex-based techniques were the next and the data-driven approaches were the least. Zero-DCE took little time for its simple curve mapping. STAR and RetinexDIP took more time due to the resolution processes during the test phases. The proposed model and retinex-based LIME took a similar amount of time. In our framework, the most time was consumed in the curve adjustment before the texture extractor, and the curve learning and texture enhancer had high efficiency. In summary, the proposed method is efficient and practical for relevant applications.

5. Conclusion

In this study, a novel low lumination image enhancement method was proposed that incorporates a transformer-based curve adjustment network and an encoder-decoder texture enhancement network with channel and spatial attention mechanisms. The method decomposed the problem into contrast enhancement and texture exploration. For contrast enhancement, the transformer-based curve learning network was introduced, which extracted features globally and produced the adjusting curve automatically. For texture enhancement, an encoder-decoder network was proposed, which explored the local information and suppressed the noise. Besides, Laplacian and TV regularization terms were introduced into the loss function for the naturalness enhancement and denoising. We have evaluated the effectiveness of our approach on the LOL and SID datasets, demonstrating its competitive performance and efficiency through comparisons with state-of-the-art techniques.

References

[1] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol.39, no.3, pp.355-368, 1987. Article (CrossRef Link)

[2] K. Zuiderveld, VIII.5. - Contrast Limited Adaptive Histogram Equalization, Graphics Gems, Academic Press, pp.474-485, 1994. Article (CrossRef Link)

[3] Y. Kim, "Contrast enhancement using brightness preserving bi-histogram equalization," *IEEE Trans. on Consum. Electron.*, vol.43, no.1, pp.1-8, 1997. Article (CrossRef Link)

[4] T. Celik and T. Tjahjadi, "Contextual and Variational Contrast Enhancement," *IEEE Transactions on Image Processing*, vol.20, no.12, pp.3431-3441, 2011. Article (CrossRef Link)

[5] C. Lee, C. Lee and C. Kim, "Contrast Enhancement Based on Layered Difference Representation of 2D Histograms," *IEEE transactions on image processing*, vol.22, no.12, pp.5372-5384, 2013. Article (CrossRef Link)

[6] D. J. Jobson, Z. Rahman and G. A. Woodell, "Properties and performance of a center/surround retinex," *IEEE Transactions on Image Processing*, vol.6, no.3, pp.451-462, 1997. Article (CrossRef Link)

[7] Z. Rahman, D. J. Jobson and G. A. Woodell, "Multi-scale retinex for color image enhancement," in *Proc. of 3rd IEEE International Conference on Image Processing*, vol.3, pp.1003-1006, 1996. Article (CrossRef Link)
[8] X. Guo, Y. Li and H. Ling, "LIME: Low-Light Image Enhancement via Illumination Map Estimation," *IEEE Transactions on Image Processing*, vol.26, no.2, pp.982-993, 2017. Article (CrossRef Link)
[9] M. Li, J. Liu, W. Yang, X. Sun, and Z. Guo, "Structure-Revealing Low-Light Image Enhancement Via Robust Retinex Model," *IEEE Transactions on Image Processing*, vol.27, no.6, pp.2828-2841, 2018. Article (CrossRef Link)
[10] J. Xu, Y. Hou, D. Ren, L. Liu, F. Zhu, M. Yu, H. Wang, and L. Shao, "STAR: A Structure and Texture Aware Retinex Model," *IEEE Transactions on Image Processing*, vol.29, pp.5022-5037, 2020. Article (CrossRef Link)
[11] Y. Gao, H. Hu, B. Li, and Q. Guo, "Naturalness Preserved Nonuniform Illumination Estimation for Image Enhancement Based on Retinex," *IEEE Transactions on Multimedia*, vol.20, no.2, pp.335-344, 2018. Article (CrossRef Link)
[12] X. Ren, W. Yang, W. Cheng, and J. Liu, "LR3M: Robust Low-Light Enhancement via Low-Rank Regularized Retinex Model," *IEEE Transactions on Image Processing*, vol.29, pp.5862-5876, 2020. Article (CrossRef Link)
[13] X. Fu, D. Zeng, Y. Huang, Y. Liao, X. Ding, and J. Paisley, "A fusion-based enhancing method for weakly illuminated images," *Signal processing*, vol.129, pp.82-96, 2016. Article (CrossRef Link)
[14] S. Wang, J. Zheng, H. Hu, and B. Li, "Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images," *IEEE Transactions on Image Processing*, vol.22, no.9, pp.3538-3548, 2013. Article (CrossRef Link)
[15] R. He, M. Guan and C. Wen, "SCENS: Simultaneous Contrast Enhancement and Noise Suppression for Low-Light Images," *IEEE Transactions on Industrial Electronics*, vol.68, no.9, pp.8687-8697, 2021. Article (CrossRef Link)
[16] X. Zhang, P. Shen, L. Luo, L. Zhang, and J. Song, "Enhancement and noise reduction of very low light level images," in *Proc. of the 21st International Conference on Pattern Recognition (ICPR2012)*, pp.2034-2037, 2012. Article (CrossRef Link)
[17] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu, "Fast efficient algorithm for enhancement of low lighting video," in *Proc. of 2011 IEEE International Conference on Multimedia and Expo*, pp.1-6, 2011. Article (CrossRef Link)
[18] K. G. Lore, A. Akintayo and S. Sarkar, "LLNet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol.61, pp.650-662, Jan. 2017. Article (CrossRef Link)
[19] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to See in the Dark," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.3291-3300, 2018. Article (CrossRef Link)
[20] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," in *Proc. of British Machine Vision Conference*, 2018. Article (CrossRef Link)
[21] W. Ren, S. Liu, L. Ma, Q. Xu, X. Xu, X. Cao, J. Du, and M. Yang, "Low-Light Image Enhancement via a Deep Hybrid Network," *IEEE Transactions on Image Processing*, vol.28, no.9, pp.4364-4375, 2019. Article (CrossRef Link)
[22] J. Deng, G. Pang, L. Wan, and Z. Yu, "Low-light Image Enhancement based on Joint Decomposition and Denoising U-Net Network," in *Proc. of 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp.883-888, 2020. Article (CrossRef Link)
[23] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep Light Enhancement without Paired Supervision," *IEEE Transactions on Image Processing*, vol.30, pp.2340-2349, 2021. Article (CrossRef Link)

[24] C. Li, C. Guo and C. C. Loy, "Learning to Enhance Low-Light Image via Zero-Reference Deep Curve Estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.44, no.8, pp.4225-4238, 2022. Article (CrossRef Link)
[25] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse Gradient Regularized Deep Retinex Network for Robust Low-Light Image Enhancement," *IEEE Transactions on Image Processing*, vol.30, pp.2072-2086, 2021. Article (CrossRef Link)
[26] Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, and F. Kuang, "RetinexDIP: A Unified Deep Framework for Low-Light Image Enhancement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.32, no.3, pp.1076-1088, 2022. Article (CrossRef Link)
[27] G. Kim, D. Kwon and J. Kwon, "Low-Lightgan: Low-Light Enhancement Via Advanced Generative Adversarial Network With Task-Driven Training," in *Proc. of 2019 IEEE International Conference on Image Processing (ICIP)*, pp.2811-2815, 2019. Article (CrossRef Link)
[28] C. Li, J. Guo, F. Porikli, and Y. Pang, "LightenNet: A Convolutional Neural Network for weakly illuminated image enhancement," *Pattern Recognition Letters*, vol.104, pp.15-22, 2018. Article (CrossRef Link)
[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. of the International Conference on Learning Representations (ICLR 2021)*, 2021. Article (CrossRef Link)
[30] E. H. Land, "The Retinex Theory of Color Vision," *Scientific American*, vol.237, no.6, pp.108-128, 1977. Article (CrossRef Link)
[31] L. Shen, Z. Yue, F. Feng, Q. Chen, S. Liu, and J. Ma, "MSR-net:Low-light Image Enhancement Using Deep Convolutional Network," *arXiv preprint arXiv:1711.02488*, 2017. Article (CrossRef Link)
[32] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proc. of 18th International Conference, Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science*, vol.9351, pp.234-241, 2015. Article (CrossRef Link)
[33] V. Lempitsky, A. Vedaldi and D. Ulyanov, "Deep Image Prior," in *Proc. of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9446-9454, 2018. Article (CrossRef Link)
[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. of 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp.6000-6010, USA, 2017. Article (CrossRef Link)
[35] S. A. Bigdeli and M. Zwicker, "Image Restoration using Autoencoding Priors," *arXiv preprint arXiv:1703.09964*, 2017. Article (CrossRef Link)
[36] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of* ICML '08: Proceedings of the 25th international conference on Machine learning, pp.1096-1103, 2008. Article (CrossRef Link)
[37] X. Mao, C. Shen and Y. Yang, "Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks with Symmetric Skip Connections," in *Proc. of 30th Conference on Neural Information Processing Systems (NIPS 2016)*, 2016. Article (CrossRef Link)
[38] W. Ren, J. Tian, Q. Wang, and Y. Tang, "Dually Connected Deraining Net Using Pixel-Wise Attention," *IEEE Signal Processing Letters*, vol.27, pp.316-320, 2020. Article (CrossRef Link)
[39] S. Yin, Y. Wang and Y. Yang, "Attentive U-recurrent encoder-decoder network for image dehazing," *Neurocomputing*, vol.437, pp.143-156, 2021. Article (CrossRef Link)
[40] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp.2536-2544, 2016. Article (CrossRef Link)
[41] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," in *Proc. of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1777-1786, 2020. Article (CrossRef Link)

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol.13, no.4, pp.600-612, 2004. Article (CrossRef Link)
[43] A. Mittal, R. Soundararajan and A. C. Bovik, "Making a "Completely Blind" Image Quality Analyzer," *IEEE Signal Processing Letters*, vol.20, no.3, pp.209-212, 2013. Article (CrossRef Link)

Yulin Cao received the B.S. degree from Qinghai Normal University, Xining, Qinghai, China, in 1994, the M.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2006, and the Ph.D. degree from the School of Computer Science, Shaanxi Normal University, Xi'an, in 2019.
He is currently a Professor with the College of Computer, Qinghai Normal University. His research expertise lies in Internet of Things, social network, and data privacy.

Chunyu Li was born in 1996. He received the B.E. degree in computer science and technology from the Jinling Institute of Technology, Nanjing, China, in 2018. He is currently perusing the M.E. degree in software engineering with Nanjing University of Information Science and Technology (NUIST), Nanjing, China. His current research interests include multimedia processing and image enhancement.

Guoqing Zhang (Member, IEEE) received the B.S. and master's degrees in information engineering from Yangzhou University, Yangzhou, China, in 2009 and 2012, respectively, and the Ph.D. degree in pattern recognition and intelligence system from the Nanjing University of Science and Technology, Nanjing, China, in 2017. He is currently an Associate Professor with the School of Computer Science, Nanjing University of Information Science and Technology (NUIST), Nanjing. His current research interests include computer vision, pattern recognition, and machine learning.

Yuhui Zheng (Member, IEEE) was born in Shanxi, China, in 1982. He received the B.S. degree in chemistry and his Ph.D. degree in computer science from Nanjing University of Science and Technology (NJUST), Nanjing, Jiangsu, China, in 2004 and 2009, respectively. From 2014 to 2015, he was a visiting professor in the digital media laboratory of the school of Electronic and Electrical Engineering, Sungkyunkwan University, Korea. He is currently a Full Professor at the School of Computer and Software in Nanjing University of Information Science and Technology (NUIST). His main research areas include image and video analysis, scene understanding, visual tracking, and pattern recognition.