BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/06 pp110–119
Volume 5, Number 2, June 2022
DOI: 10.26599/BDMA.2021.9020020

MAGAN: Unsupervised Low-Light Image Enhancement Guided by Mixed-Attention

Renjun Wang, Bin Jiang*, Chao Yang, Qiao Li, and Bolin Zhang

Abstract: Most learning-based low-light image enhancement methods typically suffer from two problems. First, they require a large amount of paired data for training, which are difficult to acquire in most cases. Second, in the process of enhancement, image noise is difficult to be removed and may even be amplified. In other words, performing denoising and illumination enhancement at the same time is difficult. As an alternative to supervised learning strategies that use a large amount of paired data, as presented in previous work, this paper presents an mixed-attention guided generative adversarial network called MAGAN for low-light image enhancement in a fully unsupervised fashion. We introduce a mixed-attention module layer, which can model the relationship between each pixel and feature of the image. In this way, our network can enhance a low-light image and remove its noise simultaneously. In addition, we conduct extensive experiments on paired and no-reference datasets to show the superiority of our method in enhancing low-light images.

Key words: low-light image enhancement; unsupervised learning; Generative Adversarial Network (GAN); mixed-attention

1 Introduction

Images captured in suboptimal lighting conditions often exhibit low contrast, unclear details, heavy noise, and low visibility. This type of image not only affects its own aesthetic quality but also is not conducive to information transmission for high-level image tasks, such as image segmentation (see Fig. 1a as an example). Such images also pose challenges in various areas of daily life, such as all-day autonomous driving, visual surveillance, and computational photography. To solve this problem, a large number of low-light image enhancement methods have been proposed.

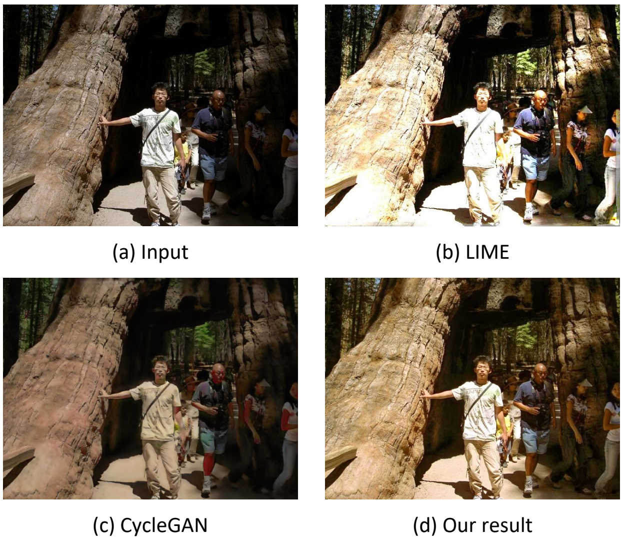

(a) Input (b) LIME (c) CycleGAN (d) Our result

Fig. 1 Example of images captured in suboptimal lighting conditions, (a) is enhanced by various methods (b)–(d). Our methods can obtain a result that contains clear details, distinct contrast, and natural color.

• Renjun Wang, Bin Jiang, Chao Yang, Qiao Li, and Bolin Zhang are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha 410006, China. E-mail: wrj@hnu.edu.cn; jiangbin@hnu.edu.cn; yangchaoedu@hnu.edu.cn; hliqiao@hnu.edu.cn; onlyou@hnu.edu.cn.

∗ To whom correspondence shoule be addressed.

Manuscript received: 2021-10-15; accepted: 2021-11-03

Low-light image enhancement methods are divided into traditional methods and learning-based methods. Traditional methods include two categories: one is based on histogram equalization[1–5], and the other is based on the retinex theory[6–12]. The former focuses on expanding the dynamic range and improving the contrast of the image, while the latter regards the image as a composition of reflection components and illumination components, treats the reflection components as the result of enhancement, and then obtains the final enhancement result by estimating the illumination component. Although they all achieved good results, both types of methods ignore the presence of image noise, causing the enhancement result to retain or even amplify noise.

Compared with traditional methods, solutions based on deep learning have higher accuracy, robustness, and speed, hence attracting increasing attention. However, most solutions, such as Refs. [13–16], ignore noise elimination in the image as well. Although some learning-based methods have addressed this issue, they adopted a simple method of connecting the enhancement module and the denoising module in series, which will cause the image to be blurred or the image noise to be magnified. Another issue is that the majority of learning-based methods[17–26] require numerous paired data for model training. However, in the field of low-light image enhancement, paired datasets are difficult to collect. In other words, collecting two corresponding images is difficult; they are the same scene and have the same content and details, and only the illumination conditions are different. Some methods, such as Refs. [19, 20], use artificial synthesis strategies to obtain low-light images to match existing normal-light images. Nevertheless, this strategy easily leads to serious artificial artifacts in generated images and poor generalization ability of the model. To address this issue, unsupervised deep learning-based methods, such as Refs. [13, 27, 28], have been proposed, yet none of them can enhance the image and denoise concurrently in a single-stage network form.

To solve the two above-mentioned major problems simultaneously, we propose a Mixed-Attention guided Generative Adversarial Network (MAGAN) to enhance a low-light image and remove its noise at the same time. Inspired by [29–32], we introduce a mixed-attention module layer to model the relationship between each pixel and features of the image and guide the network to remove image noise. To the best of our knowledge, we are the first to integrate a mixed-attention mechanism into a Generative Adversarial Network (GAN) for unsupervised low-light image enhancement. In addition, our network is lightweight and can be trained in a totally unsupervised manner. We compare our method with several latest methods to prove its superiority (see Fig. 1 for an example). The retinex model-based method, LIME[6], suffers from heavy over exposure, and the unsupervised deep learning method, CycleGAN[33], exhibits obvious color distortion. Unlike both methods, our methods can obtain a result that contains clear details, distinct contrast, and natural color.

Overall, this work's contributions are threefold:

- To overcome the difficulty of insufficient paired datasets, we propose a single-stage framework for enhancing low-light images and denoise concurrently in a fully unsupervised way.
- We introduce a mixed-attention module layer to model the relationship between each pixel and features of an image and integrate it into the network to guide the process of enhancement and denoising.
- To prove the superiority of our model, we perform extensive experiments using paired and no-reference datasets on various traditional and learning-based models qualitatively and quantitatively.

2 Related Work

Low-light image enhancement has been a subject of study for a long time. In this section, we will briefly review the methods used in this field.

2.1 Traditional low-light image enhancement methods

Traditional low-light image enhancement methods are divided into two categories: histogram equalization-based methods and retinex model-based methods.

(1) Histogram equalization-based methods

Histogram equalization generally uses histogram equalization to enhance the contrast of an image, which can improve the quality of low-light images to a certain extent. It has various implementation patterns, such as global histogram equalization, local histogram equalization, histogram specification, and dynamic histogram specification. However, some of these implementations are computationally intensive and time consuming, and some cannot change the dynamic range of the image. AHE[5] uses the histogram distribution in the local area window to construct the mapping function in the process of equalization. BPDHE[2] aims to dynamically maintain image brightness. DHECI[4] introduces a differential grayscale histogram. LDR[3] aims to expand the difference in the gray level of adjacent pixels through the layered difference representation of the 2D histogram. Nevertheless, this type of method focuses on image contrast enhancement, ignoring the impact of illumination conditions on the image. The enhanced image will exhibit color distortion, unnaturalness, overexposure/underexposure, artifacts, or noise amplification.

(2) Retinex model-based methods

These methods are based on the retinex theory. The theory treats an image as consisting of two parts – reflection component and illumination component – through some kind of prior or regularization. Then, the reflection component is approximated as the final enhancement result. Typical methods include LIME[6], NPE[11], MSR[7], and BIMEF[12]. LIME finds the maximum value in RGB channels to estimate the illumination of each pixel and then imposes a structure prior to constructing the illumination map. NPE adopts a balanced strategy to avoid overexposure. MSR restores the illumination component of the original image. BIMEF proposes a fusion algorithm for low-light image enhancement. Reference [8] proposed a robust retinex model to take image noise into account. Such methods have several limitations. For instance, the assumption that the reflection component can be approximated as the final enhancement result does not always hold under various illumination conditions, and these methods usually ignores the noise and even magnifies it.

2.2 Deep learning-based low-light image enhancement methods

Since LLNet, the first seminal deep learning-based low-light image enhancement method[17], was proposed, a large number of learning-based methods have emerged in recent years, most of which have achieved compelling success.

MBLLEN[20] decomposes the whole enhancement process into three modules — a feature extraction module, an enhancement module, and a fusion module — to build an end-to-end network. Reference [19] adds two subnetworks on the basis of MBLLEN — one is an attention network that focuses on image illumination enhancement and denoising, and the other is a reinforcement network that enhances image contrast. Reference [21] utilized an encoder-decoder network to enhance the image content and a recurrent neural network to enhance the image edge. TBEFN[18] enhances low-light images in two branches, then fuses them and implements refinement. Xu et al.[24] proposed a frequency-based decomposition-and-enhancement network to enhance an image with noise suppression. Wang et al.[16] introduced intermediate illumination based on the retinex theory to associate the input with the expected enhancement result. Reference [22] proposed a progressive retinex model, which can estimate the illumination map and noise level simultaneously. KinD[25] developed three subnetworks for layer decomposition, reflectance restoration, and illumination adjustment, separately. Fan et al.[34] borrowed ideas from image super-resolution[35,36], face restoration[37], saliency detection[38,39], text-to-image synthesis[40], and recommendation work[41], embedding semantic segmentation as prior information into the network to assist image enhancement.

However, the above-mentioned methods all require a large amount of paired data for training, which are difficult to collect in many cases. If the strategy of synthesizing low-light datasets from normal-light datasets similar to that used by Refs. [19,20] is applied, the model will have poor generalization ability, and the enhancement result will exhibit heavy artificial artifacts. Therefore, some unsupervised learning methods that do not require paired training data have been proposed. CycleGAN[33] realizes unsupervised end-to-end image translation through cycle consistency. Chen et al.[13] proposed two-way GANs without the need for the paired training data. EnlightenGAN[27] adopts a self-regularized attention map and self-feature preserving loss to realize an unsupervised low-light image enhancement method. Similar to Ref. [42], Ref. [28] built a decoupled network with two stages to implement illumination enhancement and denoising separately. However, none of these methods can enhance the image and denoise concurrently in a single-stage network form.

2.3 Attention mechanism

Reference [43] applied the attention mechanism to the field of computer vision for the first time. SENet[30] assigns different weights to different channels during the CNN operation. Reference [31] proposed a non-local operation to capture long-distance dependencies. CBAM[32] proposes a serial attention module for feed-forward convolutional neural networks. ECA-Net[44] makes lightweight improvements based on SENet. Reference [29] uses feature attention to remove image noise.

3 Method

Figure 2 shows the overall architecture of our model, which is a GAN. We adopt the encoder-decoder architecture for the generator. The mixed-attention module layer is embedded in the generator to guide image enhancement and denoising. For the discriminator, we use PatchGAN for global and local discrimination. In this section, we will illustrate our network and the mixed-attention module layer in detail.

3.1 Network architecture

Given a low-light image I, the final target is to obtain a normal-light image I^O. We adopt the joint learning form of GANs to achieve this goal. The generator G takes the encoder-decoder form. For the encoder, we perform four downsampling operations and insert the feature attention layer, which is a component of the mixed-attention module layer, in the first level. In the second to fifth levels, we embed the mixed-attention module layer. Then, the feature map undergoes four upsampling operations. To avoid the problem of information loss caused by the convolution operation in the upsampling process, all upsampling is performed using the PixelShuffle[45] operation. In the last four levels, we also insert the feature attention layer. Eventually, the feature map becomes three channels through the last convolutional layer and is then modeled in a residual manner, i.e., the original image is added element-wise to obtain the final result I^O.

$$I^O = I + G(I) \tag{1}$$

We use PatchGAN as the discriminator, which has a fully convolutional form. Similar to Ref. [27], we use both global and local discriminators. For the global discriminator, we take the entire image as input, and for the local discriminator, we take random cropped patches from both output and real normal-light images as input. The global and local discriminators consist of 7 and 6 convolutional layers, respectively. We also randomly crop patches from the original image. The enhanced image/patches and the original image/patches are sent into the pretrained VGG-16 to narrow the gap between the content. Inspired by Ref. [28], when calculating content loss, we have an instance normalization on the VGG-16 feature maps to reduce the influence of brightness on image content, which allows the network to pay more attention to the image content and stabilize training.

3.2 Loss functions

To achieve unsupervised learning for our model, we adopt the learning method of the GAN to ensure that the sample distribution of the image enhanced by our model is close to the sample distribution of the normal-light image. Otherwise, we use the content loss as well to maintain consistency between the enhanced image and the original image content.

For the global and local discriminator, we use the standard LSGAN[46] as the adversarial loss. When the discriminator is updated, the adversarial loss can be formally written as

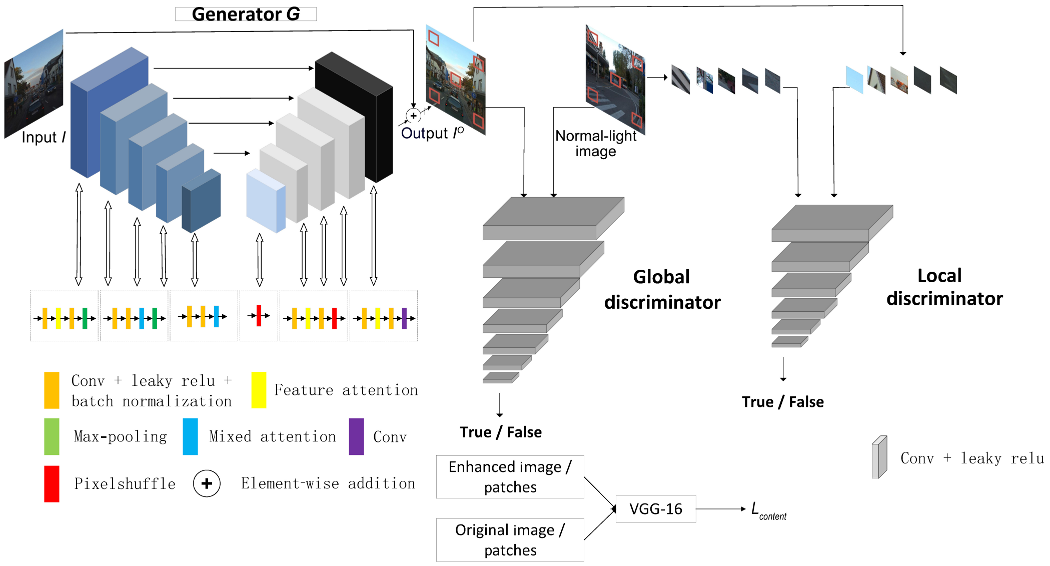


Fig. 2 Overview of MAGAN composed of one generator and two discriminators. The generator is an encoder-decoder architecture and contains our mixed-attention module layer. The global discriminator takes the entire image as input, whereas the local discriminator takes random cropped patches from both output and real normal-light image as input.

$$L_D = E_{x\sim P_{data}(x)}\left[(D(x)-1)^2\right] + E_{z\sim P_z(z)}\left[(D(G(z)))^2\right] \tag{2}$$

When updating the generator,

$$L_G = E_{z\sim P_z(z)}\left[(D(G(z)) - 1)^2\right] \tag{3}$$

where G and D denote the generator and discriminator, respectively. For the global discriminator, $P_{data}(x)$ represents the real image (normal-light image) distribution, $P_z(z)$ represents the fake image (generated image) distribution, and the adversarial losses of D an G are named as named L_D^{global} and L_G^{global}, respectively. For the local discriminator, $P_{data}(x)$ represents the real image patch (normal-light image patch) distribution, $P_z(z)$ represents the fake image patch (generated image patch) distribution, and the adversarial losses of D an G are named as named L_D^{local} and L_G^{local}, respectively. x is the samples taken from $P_{data}(x)$, and z is the samples taken from $P_z(z)$. E represents the expected value.

To ensure the perceptual quality of image content, we use the content loss to ensure consistency between the enhanced image/patch and the original image/patch. It is defined as

$$L_{content} = \frac{1}{w_{ij}h_{ij}c_{ij}}\sum_{x=1}^{w_{ij}}\sum_{y=1}^{h_{ij}}\sum_{z=1}^{c_{ij}} \| ins\left(\Phi_{ij}\left(I^O\right)\right) - ins\left(\Phi_{ij}(I)\right)\|_2^2 \tag{4}$$

where I and I^O denote the original image/patch and the enhanced image/patch, respectively. $\Phi_{ij}()$ indicates extracting the feature map obtained by the j-th convolutional layer in the i-th block of the VGG-16 network. $ins()$ denotes the instance normalization operation. w_{ij}, h_{ij}, and c_{ij} describe the dimensions of the corresponding feature map. For the whole image and the image patch, the content losses are named $L_{content}^{global}$ and $L_{content}^{local}$, respectively.

The total loss is expressed as

$$L_{total} = L_G^{global} + L_G^{local} + L_{content}^{global} + L_{content}^{local} \tag{5}$$

3.3 Mixed-attention module layer

The mixed-attention module layer in our model enhances the image features that need to be enhanced and, at the same time, removes image noise. It is composed of feature attention and pixel attention and is able to model the relationship between each pixel and features of an image. Through our mixed-attention module layer, the feature map can be refined adaptively at every layer of the network. Figure 3 illustrates the structure of the mixed-attention module layer, where H, W, and C represent the dimensions of the feature map.

Feature attention first divides the input feature map $X \in \mathbf{R}^{H\times W\times C}$ into two branches and compresses them in the spatial dimension to be $X' \in \mathbf{R}^{1\times 1\times C}$. Then, the channel dimensions of the two branches at the same multilayer perceptron are compressed to reduce network calculation complexity. After the channel dimension is recovered, the two branches are summed up, and a sigmoid function is passed to obtain the feature attention map $M_F \in \mathbf{R}^{1\times 1\times C}$. The feature attention map is then multiplied by the original map X element-wise to obtain the intermediate feature map $X^F \in \mathbf{R}^{H\times W\times C}$. Accordingly, the weights of feature attention can be computed as

$$w = \sigma\left\{W_2 ReLU\left[W_1 f(x)\right] + W_2 ReLU\left[W_1 g(x)\right]\right\} \tag{6}$$

where $f(x)$ and $g(x)$ denote the max-pooling input feature map and the average-pooling input feature map, respectively. W_1 and W_2 are set to $C \times (C/r)$ and $(C/r) \times C$, respectively, to reduce calculation complexity, where r denotes the reduction ratio. $ReLU$ and σ indicate the $ReLU$ activation function and the sigmoid function, respectively.

Then, X^F is used as the input for pixel attention, which is also divided into two branches for compression in the channel dimension to be $X^{F'} \in \mathbf{R}^{H\times W\times 1}$. Two branches of the feature map are connected after a

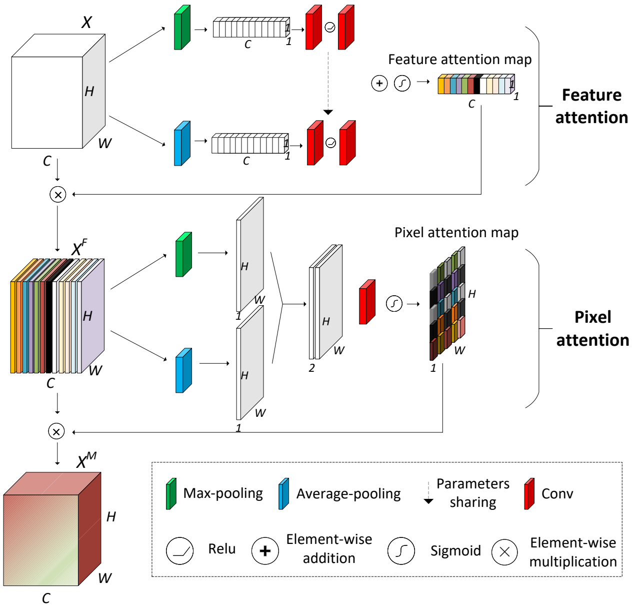


Fig. 3 Diagram of the mixed-attention module layer. It is composed of two attention module layers: the feature attention module layer and the pixel attention module layer.

convolution layer and a sigmoid function to obtain the pixel attention map $M_P \in \mathbf{R}^{H \times W \times 1}$. Finally, the pixel attention map M_P and X^F are multiplied element-wise to obtain the final refined output $X^M \in \mathbf{R}^{H \times W \times C}$. Hence, the weights of the pixel attention can be learned by

$$w = \sigma\left(conv\left(concat\left(f\left(x\right),\ g\left(x\right)\right)\right)\right) \tag{7}$$

where *concat*( ) denotes the concatenation operation; *conv*( ) denotes the convolution operation.

Feature attention can model the relationship between features of an image and enable the network to learn which features should be overenhanced and which ones should not be overenhanced, while pixel attention can model the relationship between each pixel of an image. The network can adaptively learn the weights between individual pixel points in the image, and the weights of the noise points are usually learned to a lower value. In this way, each pixel in the image can establish a connection with all other pixels, thereby highlighting commonalities and eliminating differences (usually noise).

4 Experiment

In this section, quantitative and qualitative experiments are conducted to evaluate the performance of our model. We also perform an ablation study on the mixed-attention module layer to fully validate its effectiveness on our model.

4.1 Datasets and implementation details

We trained our model on an unpaired dataset collected from Ref. [27], which contains 914 low-light images and 1016 normal-light images. To evaluate our performance, we gathered 200 no-reference low-light images from Refs. [6, 11] to test our model. We also tested our model on a dataset of 148 paired images collected from datasets, such as LOL[23], to validate its denoising ability. Low-light images in the paired dataset contain noise produced during the photo capture process.

We built our network on PyTorch and trained it for 200 epochs on a PC with NVIDIA GeForce GTX 1080 Ti GPU and 11GB memory. The network was optimized using the Adam optimizer with a learning rate of 10^{-4} for the first 100 epochs and a linear decrease to 0 for the next 100 epochs. For data augmentation, we performed a random crop to obtain a 320 × 320 patch from the raw image followed by a random horizontal flip. The batch size was set to 10 for training.

4.2 Qualitative study

We performed qualitative experiments on the no-reference low-light image dataset and paired dataset mentioned above. Our network was compared with both traditional and recent learning-based methods. Figures 4 and 5 show the representative comparison results of the no-reference and paired test datasets, respectively.

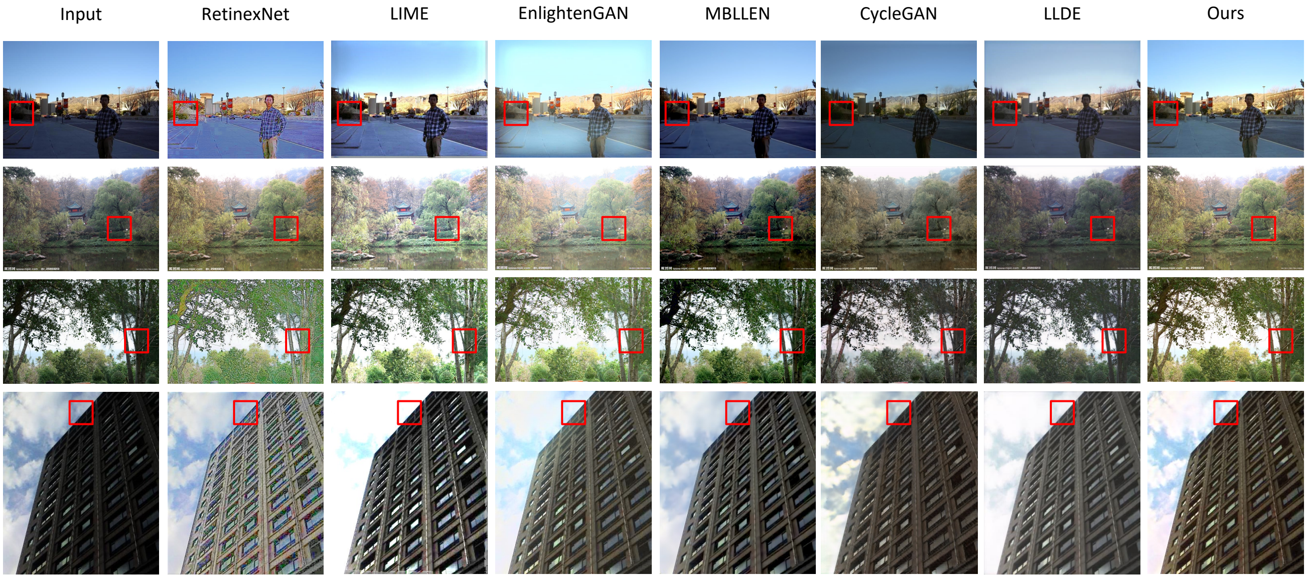


Fig. 4 Visual comparison of the results on the no-reference test dataset enhanced by RetinexNet[23], LIME[6], EnlightenGAN[27], MBLLEN[20], CycleGAN[33], LLDE[24], and MAGAN, where CycleGAN is trained using the training set from this paper. The first six methods generate images that are overexposed, underexposed, or have color deviations or unclear details. In contrast, our model generates images with clear details, accurate colors, and distinct contrast. Red boxes indicate the noisy regions where most existing methods fail.

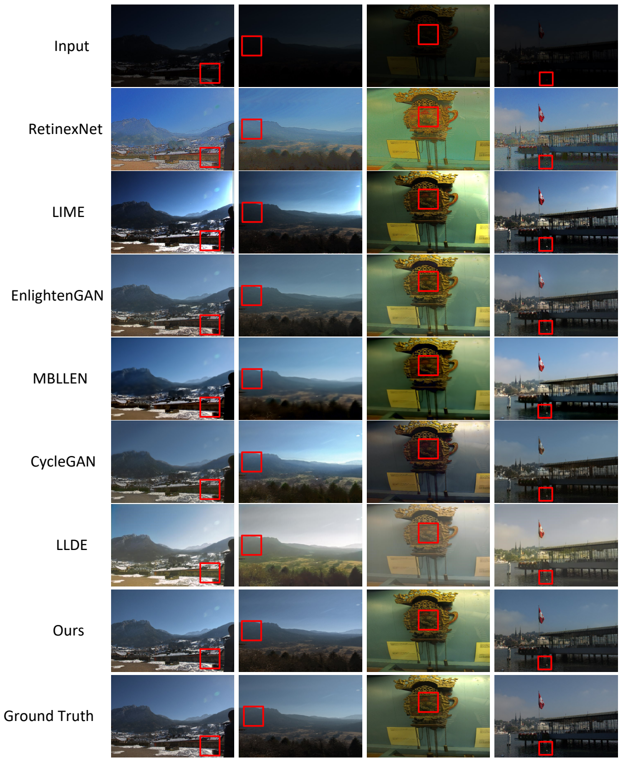


Fig. 5 Visual comparison of the results on the paired test dataset enhanced by RetinexNet[23], LIME[6], EnlightenGAN[27], MBLLEN[20], CycleGAN[33], LLDE[24], and MAGAN. The images generated by our model have the highest similarity to the ground truth and are almost free of noise. Red boxes indicate the noisy regions where most existing methods fail.

As we can see in Fig. 4, for the first row, images generated by LIME, MBLLEN, CycleGAN, and LLDE suffer from underexposure. The image generated by RetinexNet has significant color distortion, and the image generated by EnlightenGAN has obvious overexposure. For the middle two rows, the image enhanced by our model is the most natural. For the last row, LIME enhances images with loss of detail. Images enhanced by RetinexNet and CycleGAN are blurry. EnlightenGAN, MBLLEN, and LLDE enhance images with low contrast. Checking the details shows that our model achieves the best qualitative quality.

In Fig. 5, the images generated by RetinexNet and LLDE suffer from color bias and excessive noise. The contrast of the images generated by LIME and MBLLEN is too high, and the images generated by MBLLEN have excessive smoothing. The images generated by EnlightenGAN and CycleGAN are overenhanced compared with the ground truth. Contrary to all the other methods, our model has the best visual performance, with the image having the highest similarity to the ground truth and containing almost no noise.

4.3 Quantitative study

We performed a quantitative evaluation of all generated images. For the no-reference test dataset, we used a reference-free image evaluation metric Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) for evaluation because all generated images are not supervised by reference images. The overall principle of the algorithm is to extract the Mean Subtracted Contrast Normalized (MSCN) coefficients from the image, fit the MSCN coefficients to asymmetric generalized Gaussian distribution, extract the features of the fitted Gaussian distribution, and input them to a Support Vector Machine (SVM) for regression to obtain the image quality evaluation results. A small BRISQUE value corresponds to a higher representation of the image quality. For the paired test dataset, we use the PSNR and SSIM values to evaluate the model performance; large PSNR and SSIM values correspond to improved model performance. The experimental results are shown in Table 1.

Our model achieves the best performance except for BRISQUE, which is inferior to CycleGAN only. This finding shows the good superiority of our model.

4.4 Ablation study

To fully validate the effectiveness of the mixed-attention module layer on our model, we performed an ablation study. First, we removed the mixed-attention module layer from the model. Then, we removed the feature attention and pixel attention from the mixed-attention module layer in the model. Other hyperparameters were kept constant throughout the experiment. The models were trained and tested with the same datasets and metrics. Table 2 shows the quantitative results.

As shown in Table 2, the mixed-attention module layer has a strong superiority in the model.

Table 1 Average BRISQUE results on the no-reference test dataset, and average results of PSNR and SSIM on the paired test dataset.

Method	BRISQUE	PSNR	SSIM
RetinexNet	62.6888	16.4051	0.6744
LIME	56.6476	17.1019	0.7753
EnlightenGAN	43.4444	19.4903	0.8413
MBLLEN	48.5092	19.5440	0.8374
CycleGAN	**33.6521**	20.8784	0.7882
LLDE	43.1376	18.8779	0.8066
MAGAN	37.9784	**22.3895**	**0.8470**

Table 2 Ablation study.

Condition	BRISQUE	PSNR	SSIM
Without mixed-attention	44.8738	19.5343	0.8368
With pixel attention, without feature attention	38.1645	19.4078	0.8249
With feature attention, without pixel attention	40.7780	21.2273	0.8412
With mixed-attention	**37.9784**	**22.3895**	**0.8470**

We randomly selected representative comparison results from the no-reference test dataset; Fig. 6 shows the result. In the first row, except for the model with the mixed-attention module, the images generated by all the models have a large amount of noise. In the second row, only the sky in the image generated by the model with the full mixed-attention module looks natural.

5 Conclusion

This paper proposes MAGAN for enhancing low-light images and performing denoising concurrently in a fully unsupervised way. We adopt the GAN and integrate a mixed-attention module layer into our model. The mixed-attention module layer consists of two attention procedures: feature attention and pixel attention. Feature attention can model the relationship between the features of an image and enable the network to learn which features should be overenhanced and which ones should not be overenhanced, while pixel attention can model the relationship between each pixel of an image to highlight commonalities and denoise. To the best of our knowledge, we are the first to integrate a mixed-attention mechanism into a GAN for unsupervised low-light image enhancement. In addition, we conducted extensive experiments to verify the superiority of our model.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 62072169) and Changsha Science and Technology Research Plan (No. KQ2004005).

References

[1] M. Abdullah-Al-Wadud, M. H. Kabir, M. A. A. Dewan, and O. Chae, A dynamic histogram equalization for image contrast enhancement, *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 593–600, 2007.

[2] H. Ibrahim and N. S. P. Kong, Brightness preserving dynamic histogram equalization for image contrast enhancement, *IEEE Trans. Consum. Electron.*, vol. 53, no. 4, pp. 1752–1758, 2007.

[3] C. Lee, C. Lee, and C. S. Kim, Contrast enhancement based on layered difference representation of 2D histograms, *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5372–5384, 2013.

[4] K. Nakai, Y. Hoshi, and A. Taguchi, Color image contrast enhacement method based on differential intensity/saturation gray-levels histograms, in *Proc. of the 2013 Int. Symp. Intelligent Signal Processing and Communication Systems*, Naha, Japan, 2013, pp. 445–449.

[5] 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, *Comput. Vis. Graph. Image Process.*, vol. 39, no. 3, pp. 355–368, 1987.

[6] X. J. Guo, Y. Li, and H. B. Ling, LIME: Low-light image enhancement via illumination map estimation, *IEEE Trans. Image Process.*, vol. 26, no. 2, pp. 982–993, 2017.

[7] D. J. Jobson, Z. Rahman, and G. A. Woodell, A multiscale retinex for bridging the gap between color images and the human observation of scenes, *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, 1997.

[8] M. D. Li, J. Y. Liu, W. H. Yang, X. Y. Sun, and Z. M. Guo, Structure-revealing low-light image enhancement via robust retinex model, *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2828–2841, 2018.

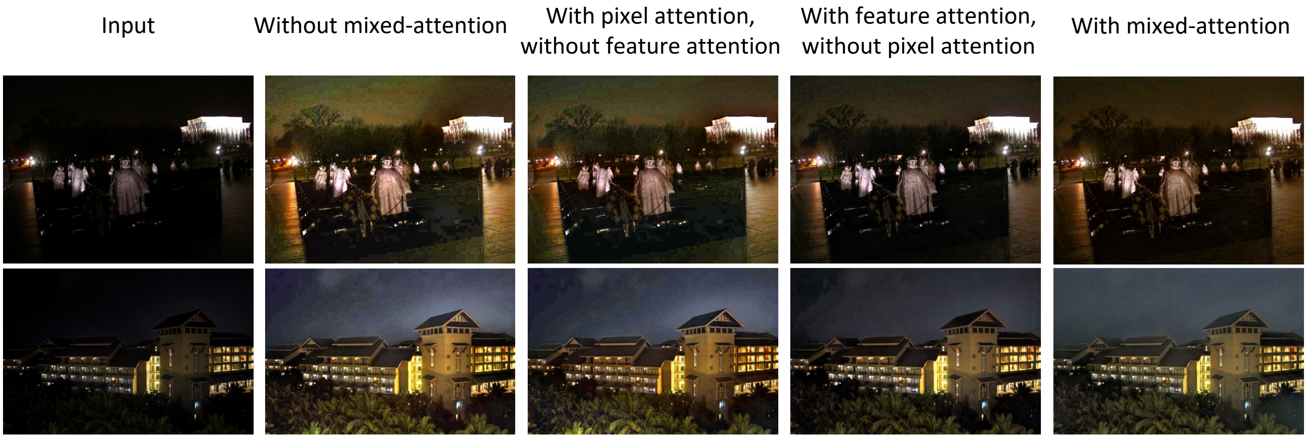

Input
Without mixed-attention
With pixel attention, without feature attention
With feature attention, without pixel attention
With mixed-attention

Fig. 6 Ablation study. Visual comparison of the results on the no-reference test dataset under each condition. In the absence of any of the attention modules, the images generated by the model are noisy and unnatural.

[9] S. Park, S. Yu, B. Moon, S. Ko, and J. Paik, Low-light image enhancement using variational optimization-based retinex model, *IEEE Trans. Consum. Electron.*, vol. 63, no. 2, pp. 178–184, 2017.

[10] X. T. Ren, W. H. Yang, W. H. Cheng, and J. Y. Liu, LR3M: Robust low-light enhancement via low-rank regularized retinex model, *IEEE Trans. Image Process.*, vol. 29, pp. 5862–5876, 2020.

[11] S. H. Wang, J. Zheng, H. M. Hu, and B. Li, Naturalness preserved enhancement algorithm for non-uniform illumination images, *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3538–3548, 2013.

[12] Z. Q. Ying, G. Li, and W. Gao, A bio-inspired multi-exposure fusion framework for low-light image enhancement, arXiv preprint arXiv: 1711.00591, 2017.

[13] Y. S. Chen, Y. C. Wang, M. H. Kao, and Y. Y. Chuang, Deep photo enhancer: Unpaired learning for image enhancement from photographs with GANs, in *Proc. of the the 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 6306–6314.

[14] A. Ignatov, N. Kobyshev, R. Timofte, and K. Vanhoey, DSLR-quality photos on mobile devices with deep convolutional networks, in *Proc. of the 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 3297–3305.

[15] A. Ignatov, N. Kobyshev, R. Timofte, K. Vanhoey, and L. Van Gool, WESPE: Weakly supervised photo enhancer for digital cameras, in *Proc. of the 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops*, Salt Lake City, UT, USA, 2018, pp. 691–700.

[16] R. X. Wang, Q. Zhang, C. W. Fu, X. Y. Shen, W. S. Zheng, and J. Y. Jia, Underexposed photo enhancement using deep illumination estimation, in *Proc. of the 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019, pp. 6842–6850.

[17] K. G. Lore, A. Akintayo, and S. Sarkar, LLNet: A deep autoencoder approach to natural low-light image enhancement, *Pattern Recogn.*, vol. 61, pp. 650–662, 2017.

[18] K. Lu and L. H. Zhang, TBEFN: A two-branch exposure-fusion network for low-light image enhancement, *IEEE Trans. Multimed.*, vol. 23, pp. 4093–4105, 2020.

[19] F. F. Lv, Y. Li, and F. Lu, Attention-guided low-light image enhancement, arXiv preprint arXiv: 1908.00682, 2019.

[20] F. F. Lv, F. Lu, J. H. Wu, and C. Lim, MBLLEN: Low-light image/video enhancement using CNNs, in *British Machine Vision Conf.* (*BMVC*), Northumbria, UK, 2018, p. 220.

[21] W. Q. Ren, S. F. Liu, L. Ma, Q. Q. Xu, X. Y. Xu, X. C. Cao, J. P. Du, and M. H. Yang, Low-light image enhancement via a deep hybrid network, *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4364–4375, 2019.

[22] Y. Wang, Y. Cao, Z. J. Zha, J. Zhang, Z. W. Xiong, W. Zhang, and F. Wu, Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement, in *Proc. 27^{th} ACM Int. Conf. Multimedia*, Nice, France, 2019, pp. 2015–2023.

[23] C. Wei, W. J. Wang, W. H. Yang, and J. Y. Liu, Deep retinex decomposition for low-light enhancement, arXiv preprint arXiv: 1808.04560, 2018.

[24] K. Xu, X. Yang, B. C. Yin, and R. W. H. Lau, Learning to restore low-light images via decomposition-and-enhancement, in *Proc. of the 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020, pp. 2278–2287.

[25] Y. H. Zhang, J. W. Zhang, and X. J. Guo, Kindling the darkness: A practical low-light image enhancer, in *Proc. 27^{th} ACM Int. Conf. Multimedia*, Nice, France, 2019, pp. 1632–1640.

[26] M. F. Zhu, P. B. Pan, W. Chen, and Y. Yang, EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network, *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 13106–13113, 2020.

[27] Y. F. Jiang, X. Y. Gong, D. Liu, Y. Cheng, C. Fang, X. H. Shen, J. C. Yang, P. Zhou, and Z. Y. Wang, EnlightenGAN: Deep light enhancement without paired supervision, *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021.

[28] W. Xiong, D. Liu, X. H. Shen, C. Fang, and J. B. Luo, Unsupervised real-world low-light image enhancement with decoupled networks, arXiv preprint arXiv: 2005.02818, 2020.

[29] S. Anwar and N. Barnes, Real image denoising with feature attention, in *Proc. of the 2019 IEEE/CVF Int. Conf. Computer Vision*, Seoul, Republic of Korea, 2019, pp. 3155–3164.

[30] J. Hu, L. Shen, and G. Sun, Squeeze-and-excitation networks, in *Proc. of the 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7132–7141.

[31] X. L. Wang, R. Girshick, A. Gupta, and K. M. He, Non-local neural networks, in *Proc. of the 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 7794–7803.

[32] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, CBAM: Convolutional block attention module, in *Proc. 15^{th} European Conf. Computer Vision*, Munich, Germany, 2018, pp. 3–19.

[33] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in *Proc. of the 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 2242–2251.

[34] M. H. Fan, W. J. Wang, W. H. Yang, and J. Y. Liu, Integrating semantic segmentation and retinex model for low-light image enhancement, in *Proc. 28^{th} ACM Int. Conf. Multimedia*, Seattle, WA, USA, 2020, pp. 2317–2325.

[35] K. C. K. Chan, X. T. Wang, X. Y. Xu, J. W. Gu, and C. C. Loy, GLEAN: Generative latent bank for large-factor image super-resolution, in *Proc. of the 2021 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Nashville, TN, USA, 2021, pp. 14240–14249.

[36] X. T. Wang, K. Yu, C. Dong, and C. C. Loy, Recovering realistic texture in image super-resolution by deep spatial feature transform, in *Proc. of the 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 606–615.

[37] X. M. Li, C. F. Chen, S. C. Zhou, X. H. Lin, W. M. Zuo, and L. Zhang, Blind face restoration via deep multi-scale component dictionaries, in *Proc. 16^{th} European Conf. Computer Vision*, Glasgow, UK, 2020, pp. 399–415.

[38] X. Lin, Z. J. Wang, L. Z. Ma, and X. B. Wu, Saliency detection via multi-scale global cues, *IEEE Trans. Multimed.*, vol. 21, no. 7, pp. 1646–1659, 2019.

[39] Z. J. Wang, L. Z. Ma, X. Lin, and H. Zhong, Saliency detection via multi-center convex hull prior, in *Proc. of the 2018 IEEE Int. Conf. Acoustics, Speech and Signal Processing* (*ICASSP*), Calgary, Canada, 2018, pp. 1867–1871.

[40] Z. X. Wang, Z. Quan, Z. J. Wang, X. J. Hu, and Y. Y. Chen, Text to image synthesis with bidirectional generative adversarial network, in *Proc. of the 2020 IEEE Int. Conf. Multimedia and Expo* (*ICME*), London, UK, 2020, pp. 1–6.

[41] W. Liu, Z. J. Wang, B. Yao, and J. Yin, Geo-ALM: Poi recommendation by fusing geographical information and adversarial learning mechanism, in *Proc. 28th Int. Joint Conf. Artificial Intelligence*, Macao, China, 2019, pp. 1807–1813.

[42] X. Lin, L. Z. Ma, B. Sheng, Z. J. Wang, and W. S. Chen, Utilizing two-phase processing with FBLS for single image deraining, *IEEE Trans. Multimed.*, vol. 23, pp. 664–676, 2020.

[43] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, Recurrent models of visual attention, arXiv preprint arXiv: 1406.6247, 2014.

[44] Q. L. Wang, B. G. Wu, P. F. Zhu, P. H. Li, W. M. Zuo, and Q. H. Hu, ECA-Net: Efficient channel attention for deep convolutional neural networks, in *Proc. of the 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, 2020, pp. 11531–11539.

[45] W. Z. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. H. Wang, Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in *Proc. of the 2016 IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 1874–1883.

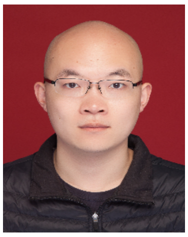
[46] X. D. Mao, Q. Li, H. R. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, Least squares generative adversarial networks, in *Proc. of the 2017 IEEE Int. Conf. Computer Vision*, Venice, Italy, 2017, pp. 2813–2821.

Renjun Wang received the BEng degree in computer science and technology from Hunan Normal University, China in 2020. He is currently a master student at the College of Computer Science and Electronic Engineering, Hunan University, China. His research interests include image enhancement, deep learning, and machine learning.

Bin Jiang received the BS and MEng degrees from Hunan University, China in 1993 and 2006, respectively, and the PhD degree from Tokyo Institute of Technology, Japan in 2015. Since 2021, he has been a professor at the College of Computer Science and Electronic Engineering, Hunan University, China. He is a member of CCF, CAAI, and ACM. His research interests include artificial intelligence, big data technology, computer vision, and machine learning.

Chao Yang received the BEng and MEng degrees in computer science from Hunan University, China in 1999 and 2005, respectively, and the PhD degree in computational intelligence and systems science from Tokyo Institute of Technology, Japan in 2010. She has ever worked as a postdoctoral researcher at Tokyo Institute of Technology. Since 2016, she has been an associate professor at the College of Computer Science and Electronic Engineering, Hunan University. She is a member of ACM and CCF. Her research interests include natural language understanding, information retrieval and recommendation, and multimedia retrieval and content understanding.

Qiao Li received the BEng degree from Wuhan University of Technology in 2004 and the MEng degree in software engineering from Central South University, China in 2010. He is currently a PhD candidate at the College of Computer Science and Electronic Engineering, Hunan University, China. His research interests include image enhancement, deep learning, and machine learning.



Bolin Zhang received the BEng degree from Henan Polytechnic University, China in 2017, and the MEng degree from Ningbo University, China in 2020. He is currently a PhD candidate at the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His current research interests include deep learning and multimodal retrieval.