IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 11, 2020, accepted May 14, 2020, date of publication May 20, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995871

Single Image Reflection Removal via Attention Model and SN-GAN

KUANHONG CHENG[1], JIANGLUQI SONG[1], JUAN DU[1], SHENGHUI RONG[2], AND HUIXIN ZHOU[1], (Member, IEEE)

[1]School of Physics and Optoelectronics Engineering, Xidian University, Xi'an 710071, China
[2]College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China

Corresponding author: Jiangluqi Song (jlqsong@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 51801142, in part by the Natural Science Foundation of Shaanxi Province under Grant 2018JQ5022, in part by the China Scholarship Council under Grant 201806960036, in part by the China Postdoctoral Science Foundation under Grant 2019M652472, and in part by the Fundamental Research Funds for the Central Universities under Grant JB180502.

ABSTRACT Single image reflection removal is of great practical importance for various computer vision tasks. Most non-learning methods try to solve this problem through the model-optimization scheme, which fails to produce promising results due to the shortage of suitable priors to model the difference between the reflection layer and the transmission layer. This paper presents an improved generative adversarial network to resolve this problem. First, we suggest that reflection removal is not only a channel-wise separation problem, but also a spatial variational occlusion removal task, which is sensitive to both spatial and channel-wise features. To this end, we integrate the CBAM module into the generator to enhance both spatial and channel-wise feature representation. Second, we consider the reflection layer as a spatial mask with space-relevant reflection intensity information, which can be used to elevate the performance of the discriminator. We then design a novel SNGAN structure with utilize the predicted reflection as a guidance to achieve better adversarial supervision. Specifically, our new generative network has an encoder-decoder structure with skip-connections, where the attention enhancement block is integrated into each skip-connection of the encoder-decoder subnet, and followed by an eight-layer fully convolutional subnet. Furthermore, the SNGAN loss is combined with L2 pixel loss and L1 VGG19 perceptual loss for training. The experimental results with benchmark datasets indicate that our method outperforms several state-of-the-art networks.

INDEX TERMS Inverse problems, artificial neural networks, image processing, image restoration, computer vision, artificial intelligence, supervised learning, multi-layer neural network, knowledge-based systems.

I. INTRODUCTION

Photos taken through glass windows usually have the undesired reflection objects overlaid on the transmission images, which will heavily reduce the image quality. Feeding those images into computer vision tasks may also degenerate the performance of entire systems. Therefore, reflection removal has recently become one of the research hotspots in the image restoration field.

Reflection removal tries to recover a clear transmission image from the mixture of transmission and reflection layers. This process can be modelled as

$$I = T + R \tag{1}$$

where T means transmission image which we want to recover, R represents the reflection layer that is to be removed. I indicates the mixture image.

Traditional model-based methods try to utilize specific priors to recover an optimized T from a given I. But it is a challenging task to find a universal prior to model the difference between T and R, as such, model-based methods cannot gain satisfactory performance. Therefore, more and more studies began to focus on CNN-based end-to-end models [13]–[17], [38], [39], [41].

The initial CNN aims to achieve different tasks through layer-by-layer convolutional scheme, where the different positions and channels of feature maps are treated equally. This is suitable for some applications like object detection or segmentation, where we want to find an integrated description for the objects. But it is not rational for reflection

The associate editor coordinating the review of this manuscript and approving it for publication was Ivan Lee.

removal because the task-relevant information is usually not uniform.

Reflection removal can be considered as a typical image separation problem, therefore, it is naturally to assume that the two layers are corresponding to some specific channels of feature maps, which means the performance of the model is sensitive to its channel-wise feature representation capability. To this end, channel-wise attention model can be used to elevate the performance. Some works [16] have applied this strategy [27] to improve the CNN model.

However, we consider that the involvement of channel attention is still not essential enough for the removal task. We hold that the reflection can also be regarded as a translucent soft mask overlapped on the transmission image, and the spatial intensity distribution of this mask is non-uniform because the colour difference and reflection strength varies between different areas of the two layers. To better illustrate this spatial non-uniformity, we integrate an attention module into our network which can enhance both spatial and channel-wise feature simultaneously.

To further dig out the non-uniformity of the spatial distribution, we also design a novel discriminator which can distinguish the difference of the reflection automatically in a learnable way. We feed the discriminator with both predicted transmission and reflection, the reflection is expected to act as a feature mask to provide more clues of the spatial non-uniformity and facilitates the discrimination. Additionally, we also apply the spectral-normalized GAN (SNGAN) to improve the stability of the discriminator. The performance of the generator in this way can also be promoted through adversarial supervision.

The contributions of this paper can be summarized as follows.

(1) We revisit the non-uniformity of the feature distribution in CNN models and propose a novel GAN-based reflection removal network, which integrates both spatial and channel-wise attention.
(2) We design a spectral-normalized discriminator which takes both the reflection and transmission layers as input, and can utilize the reflective signal in a more essential way.
(3) The performance of our proposed network outperforms existing reflection removal methods on benchmark datasets.

II. RELATED WORK

Reflection removal is a highly ill-posed problem since there is an infinite number of possible solutions for a certain input I without prior knowledge. One commonly used scheme to reduce the ill-posedness is to extract more information through image sequence [1], [2] (i.e. take multi images from different viewpoints instead of a single image as input). However, this kind of method is not suitable for practical applications since it is hard to satisfy the requirements of data acquisition in most scenarios. As a result, single image reflection removal has gained more attention in recent researches.

A. MODEL BASED METHODS

Most of the traditional single image reflection removal methods are model-optimization-based. The ill-posedness is overcome by imposing specific priors on transmission and reflection layers. It has been proved that the gradient and local features of images are sparse and have been used in varies image restoration algorithms [3]–[5]. Levin *et al.* applied this prior to reflection removal, where they try to decompose one image into two outputs by minimizing the total amount of edges and comers [6]. But this model cannot get satisfactory results when the texture of the input image is very complex. An improved method is to label the gradients for background and reflection with user assistance [7], [8], which can get better performance yet is less practical because of the requirement of manual operation. In addition, the reflection layer always suffers from blurring because it is hard to assume both the transmission layer and reflection layer are in the same focal plane, which makes the gradient prior not particularly helpful to all images. However, the different smoothness level of the two layers can also be considered as useful prior knowledge. Chung *et al.* [9] assume that even with reflection interference, most images can still focus on the transmission layer, which leads to different sharpness of the two layers, and can be separated through classifying gradient profile sharpness [10], or even by segmenting the depth of field (DoF) inspired by smooth level [11]. In addition to the above mathematic priors, some researchers attempt to achieve reflection removal through the physical model of reflective imaging systems, where they consider the thickness of the glass is non-negligible. As such, both sides of the glass can contribute to the reflection. Such ghosting effects make the reflection layer seem like an overlap of two similar reflection images with a tiny shift. Based on this assumption, Shih *et al.* [12] model the ghosted reflection by a double-impulse convolution kernel, and separate the layers using a Gaussian mixture model (GMM) as a regularization term. However, this ghost cue is not appropriate for all circumstances since the reflection of the second surface is too weak to be well observed in most scenarios. And theoretically, the number of reflection images is infinite because the reflection layer would be reflected again and again between the two surfaces, which makes this model difficult to describe real images.

B. DEEP LEARNING METHODS

Since it remains a great challenge to find a certain prior to well describe the properties of different layers, most traditional model-based methods cannot get remarkable results. Recent works try to resolve this problem with deep learning. Fan *et al.* proposed a deep architecture CEILNet [13] which is consists of two cascaded subnets with the same structure. The first subnet is used to get the target gradient, whose result is then taken as the input of the second subnet to predict the final transmission layer. Wan *et al.* introduced a concurrent

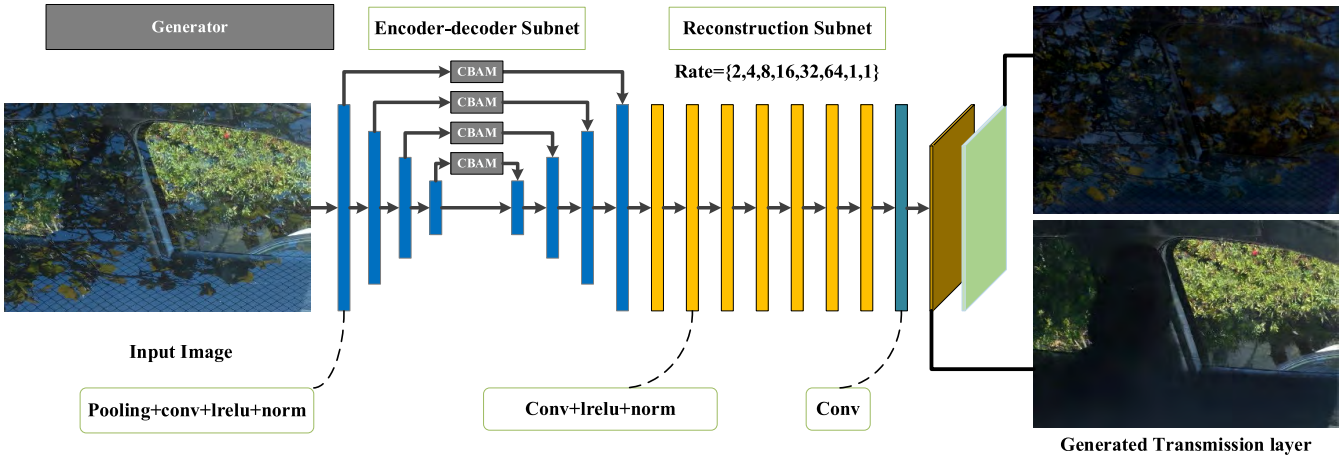


FIGURE 1. Structure of the proposed generator.

network CRRN [14], which is also a two-subnets structure and considers the gradient as an independent input. But unlike the two-stage cascaded design in CEILNet, the prediction of the target gradient and transmission layer in CRRN are carried out simultaneously, with multi-scale guided inference to connect these two subnets. Afterwards, to further remove the strong gradients left in the results, an improved method [38] is proposed to add the context information and the multi-scale gradient information into the design. Furthermore, Zhang *et al.* developed a generative adversarial network (GAN) [15]. Instead of taking the gradient as part of the input like in CEILNet and CRRN, they designed an exclusion loss to minimize the correlation between the predicted transmission and reflection layers in the gradient domain to make the network more compact. Wei *et al.* improved this model by introducing a channel-wise multi-scale attention model [16] into the decoding module of the generative network, through which the feature of different channels can be enhanced and recombined to achieve better representation. Yang [39] tried to refine the transmission by the predicted reflection through a three-stage structure, which can use the reflection layer in a more informative way. Wen *et al.* [17] first proposed a non-linear model to synthesize reflection images, which can also generate corresponding blending masks. Then a GAN architecture with three decoding modules is described to predict transmission layer, reflection layer and blending mask respectively. The loss functions are designed to minimize the error between each prediction and ground truth, as well as the input image and reconstructed reflection image synthesized through the three predictions, which make the model very comprehensive.

The above CNN methods have achieved great performance improvement compared to traditional model-optimization methods. But only the ERRNet concerns about the non-uniformity of feature maps, where they applied the channel-wise attention model SENet into the structure, and not so many works involve spatial attention model explicitly. Besides, the existing GAN-based methods only take the predicted and labelled transmission as inputs, which is also limited because the reflection signal has not been well exploited. This paper tries to solve this problem by exploiting the non-uniformity of the reflection-relevant information in both generative and discriminative networks via attention model and SN-GAN with reflection mask.

III. PROPOSED METHOD

In this section, we present the structure of our reflection removal network, as well as the loss function and training details.

A. NETWORK STRUCTURE

1) GENERATOR

The detail of this structure is shown in Figure 1. Our generative network consists of two subparts: an encoder-decoder subnet and a fully convolutional subnet (FCN).

The encoder-decoder subnet is a 4-scale 8-layer network with skip connections [18]–[20], [37]. The channel numbers of this module are set to {64, 128, 256, 512, 512, 256, 128, 64}, with filter size 3×3, each convolution is followed by an LRelu activation and normalization. We also integrate the Convolutional Block Attention Module (CBAM) [28] into the skip connection branches to enhance feature representation, this module is added to each skip connection between same-scale feature maps.

CBAM is used to enhance both channel-wise and spatial-wise feature representation. The structure is shown in Fig. 2.

First, channel-wise max-pooling and average-pooling are performed on the original feature maps to extract two vectors, which are then taken as the input of a 3-layer fully-connection net to generate channel-wise weighted values, and the enhancement is achieved by applying these values onto different feature channels. The spatial- representation is subsequently conducted in a similar way by replacing the channel-wise pooling by pixel-wise pooling and using

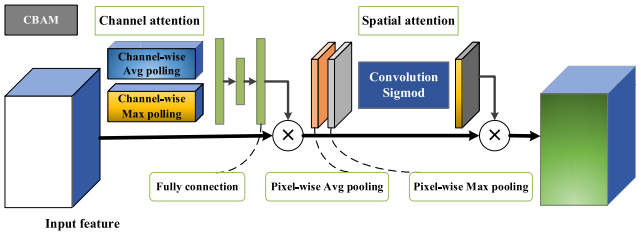


FIGURE 2. Structure of the CBAM.

the convolution layer and sigmoid activation instead of the fully-connection layer to generate a weighted map, which is then considered as enhancement coefficients to achieve the spatial representation.

Our FCN subnet has 8-layers, the channel number of the first 7 layers is 64. We use dilated convolution [22] to enhance the receptive field, the rates of these 8 dilated layers are set to {2, 4, 8, 16, 32, 64, 1, 1} with filter size 3 × 3, and same activation and normalization as the encoder-decoder subnet except the last layer. The output of the last layer has 6 channels, which is then separated into two RGB images to get the reflection and transmission predictions as in [15].

2) DISCRIMINATOR

The discriminator in our model is a 5-layer network with SNGAN loss, the 5 layers have the same filter size 5 × 5 with channel number {64, 128, 256, 256, 256} and stride 2, followed by LRelu without normalization. Since we expect to use the predicted reflection as an important clue to guide the training of the discriminator, some transformations must be done to format the inputs and get the SN loss. The structure of our discriminator is shown in Fig. 3.

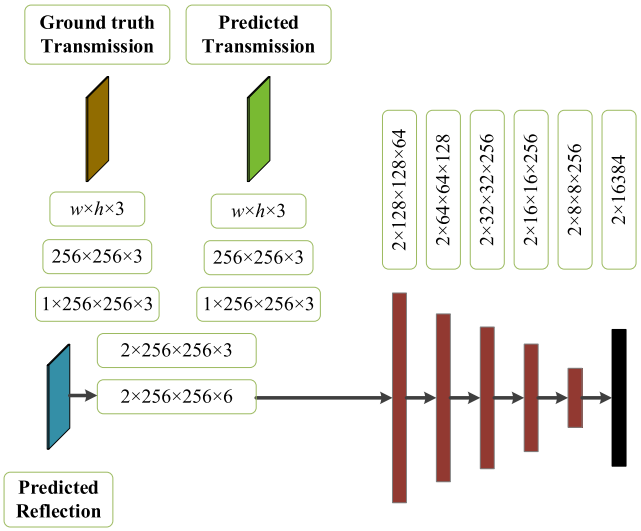


FIGURE 3. Structure of the proposed discriminator.

In our experiment, the predicted transmission and label are first resized to a specific size, that is [256, 256, 3]. Then, they are extended into two tensors (i.e. [1, 256, 256, 3]) and concatenated through axis 0 to form a two-batch four-dimensional array (i.e. [2, 256, 256, 3]). After that, we append the predicted reflection to this array along axis 3 to get a [2, 256, 256, 6] matrix, which is finally treated as the input of the discriminator. The reason we append reflection twice (i. e. at the channel axis of the two batches) is that our final output is flattened along the batch axis, so the guidance mask cannot be shared across different batches during the training.

After 5 layers of convolution, the output shall be a tensor with a size of [2, 8, 8, 256], then it is flattened into two vectors with 16384 elements to generate the generative loss and discriminative loss.

The most used DCGAN can achieve high PSNR results, but may also cause mode collapse, thus we apply the SNGAN [40] with hinge loss in our model. The loss functions of SNGAN are

$$L_D = E\left(ReLU(1 + D^{sn}(G(I)) + ReLU(1 - D^{sn}(T)\right) \quad (2)$$

$$L_A = -E(D^{sn}(G(I))) \quad (3)$$

where L_D and L_A are discriminative loss and generative loss, E represents mean value, D^{sn} means SN discriminator, $D^{sn}(x)$ is the discriminative output of sample x (i.e. the output vector in Fig. 3). $G(I)$ is the generated reflection-free image, T indicates ground truth transmission. We use the default fast approximation algorithm of spectral normalization described in SNGAN [40].

B. OBJECTIVE FUNCTION

The objective function in our method consists of two parts: pixel-wise loss and perceptual loss.

Pixel-wise loss aims to minimize the error between the generated transmission layer and ground truth, which is a widely used loss function in image restoration problems [34], [35]. We also use this loss in our method, the specific format in our network is

$$L_{pixelR} = \lambda_1 \left\| T - \hat{T} \right\|_2 + \lambda_2 \left\| \nabla T - \nabla \hat{T} \right\|_2 \quad (4)$$

where L_{pixelR} is the pixel loss for real training images, T and $\hat{T}$ are the ground truth and prediction of the transmission layer, ∇ means gradient, $\|\|_2$ indicates L2 norm and λ is weighted value. In our experiment, we set $\lambda_1 = 0.2$ and $\lambda_2 = 0.4$.

For those simulated images, we also utilize the loss of the reflection layer, the pixel loss thus is written as

$$L_{pixelS} = \lambda_1 \left\| T - \hat{T} \right\|_2 + \lambda_2 \left\| \nabla T - \nabla \hat{T} \right\|_2 + \lambda_1 \left\| R - \hat{R} \right\|_2 + \lambda_2 \left\| \nabla R - \nabla R \right\|_2 \quad (5)$$

where L_{pixelS} is pixel-wise loss, R and $\hat{R}$ are the ground truth and prediction of the reflection layer.

It has been proved that the pixel-wise loss can achieve excellent performance in terms of PSNR but would also introduce some artifacts. Thus, we also use the perceptual loss of VGG19 as in other image restoration networks [36]. We use

the "conv1_2", "conv2_2", "conv3_2", "conv4_2", and "conv5_2" layers of VGG-19 net, which is similar to [15].

$$L_{perceptual} = \sum_l \lambda_l \left\| \Phi_l(T) - \Phi_l(\hat{T}) \right\|_1 \quad (6)$$

where Φ_l is the feature map of the lth layer of VGG19, λ_l ($l = 1, 2, 3, 4, 5$) are the weighted values of different layers which are set as {1/2.6, 1/4.8, 1/3.7, 1/5.6, 1/0.15}, $\|\|_1$ indicates L1 norm.

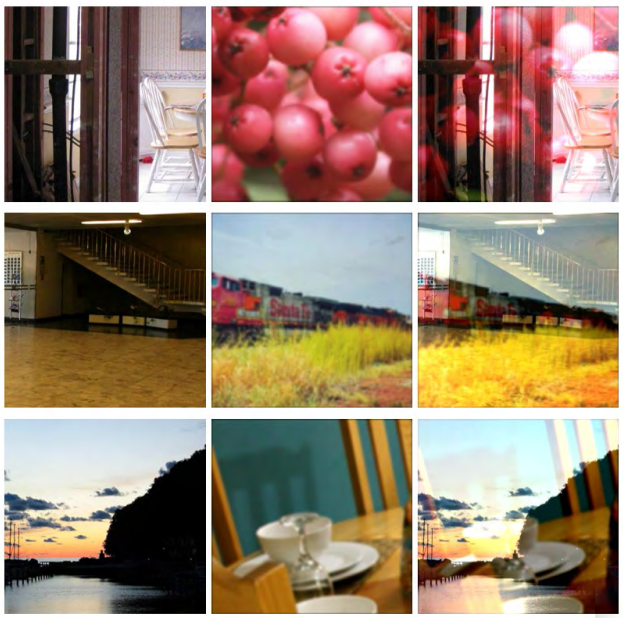
The final objective function of the generative network is set as

$$L = \alpha L_A + \beta L_{pixel} + \gamma L_{perceptual} \quad (7)$$

where α, β, and γ are weighted values and we set $\alpha = 1$, $\beta = 1$, and $\gamma = 0.5$.

C. TRAINING DETAILS

Our training dataset comes from UC Berkeley [15], which consists of two parts: the simulated data and real data. The real training data contains 90 real samples with ground truth. The simulated data includes 13700 pairs of reflection and transmission images. Our synthesis model is a weighted summary of the transmission image and smoothed reflection image as in ERRNet. We set the kernel size of gaussian filter as 11, with sigma varies from 1 to 1.5 and gamma from 2 to 5. Some of the simulated samples are shown in Fig. 4.



(a) (b) (c)

FIGURE 4. Synthesized training images. (a) transmission images, (b) reflection images, (c) synthesized samples.

Our model is trained on an RTX Titan V GPU with TensorFlow 1.9.0 for 90 epochs through an alternative optimize between the generator and discriminator. The learning rate is set to 1e-4, 3e-5, and 1e-5 for every 30 epochs. The inputs and labels are cropped from the training set with a random size varies from 200 × 200 to 384 × 384.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The performance of our proposed method is tested on 2 datasets: the real test set of UC Berkeley dataset [15], and SIR2 benchmark dataset from the ROSE lab [33]. The Berkeley set contains 20 real images with ground truth. SIR2 consists of three groups of images captured from different scenes named Wild (57 samples), Solidobject (200 samples), and Postcard (199 samples). We compare our model with CEILNet [13], Zhang *et al.* [15], and ERRNet [16] on these datasets. Since the image resolution of Berkeley dataset is too high to perform inference on TITAN V, we crop these 20 images into 87 local patches with size 500 × 500 and all the PSNR and SSIM are evaluated on these patches, we also consider the SIR2 dataset as three independent sub-datasets due to 3 different scenes as in [16].

We first make a theoretical analysis of these 4 networks in our experiment, and the attributes of these models are presented in Table 1. We consider CBAM as a 4-layer structure because it involves one convolutional layer and 3 fully-connection layers.

TABLE 1. Comparison to state of the art methods.

Model	Depth	Attention model	Perceptual loss	GAN loss
CEILNet	64	×	×	×
Zhang's	10	×	√(VGG19)	√(DCGAN)
ERRNet	32	√(SENet)	√(VGG19)	√(DCGAN)
Proposed	20	√(CBAM)	√(VGG19)	√(SNGAN)

As one of the earliest CNN-based reflection removal models, CEILNet has the deepest structure with 64 convolution layers but doesn't involve any kind of attention model, perceptual loss, or adversarial learning. Zhang's network is simple with respect to depth, but it takes the00 pre-trained VGG19 model to extract features, which can provide more useful information than the traditional layer-by-layer extraction scheme. Additionally, this model also takes GCGAN to guide the generator which may also contribute some improvements. The same strategies have also been applied to ERRNet, with a channel-wise multi-scale attention enhancement module integrated. Unlike Zhang's work and ERRNet which use DCGAN to supervise the training of the generator, we apply SNGAN to avoid mode collapse and expect to achieve better performance. Furthermore, our model involves both channel and spatial attention enhancement which is better than the channel-wise SENet used in ERRNet. Therefore, the proposed model is more comprehensive in theory.

To validate the effectiveness of the spatial-attention module in CBAM and the guidance mask in the SNGAN, we add three ablation models into the comparison process: we remove the guidance mask (i.e. remove the predicted reflection appended along axis 3 in Fig. 3, as such, the input is a tensor of [2, 2556, 256, 3]) from the discriminator, the spatial attention stage of the CBAM and the entire CBAM

TABLE 2. PSNR/SSIM on benchmark datasets.

Dataset	Metrics	Input	CEILNet	Zhang's	ERRNet	CB-N	SP-N	SN-N	Proposed
Real	PSNR	19.809	17.971	20.278	20.982	19.976	20.062	21.316	21.865
	SSIM	0.765	0.656	0.765	0.782	0.695	0.745	0.779	0.789
Wild	PSNR	26.277	20.840	21.338	23.996	21.015	22.986	23.667	25.277
	SSIM	0.899	0.808	0.833	0.853	0.697	0.839	0.868	0.886
Soild Object	PSNR	23.759	23.082	22.145	24.834	20.218	24.372	24.295	24.891
	SSIM	0.882	0.873	0.875	0.898	0.678	0.877	0.887	0.893
Postcard	PSNR	20.940	19.984	15.802	21.811	16.767	20.610	21.032	23.942
	SSIM	0.877	0.827	0.797	0.874	0.615	0.816	0.850	0.881

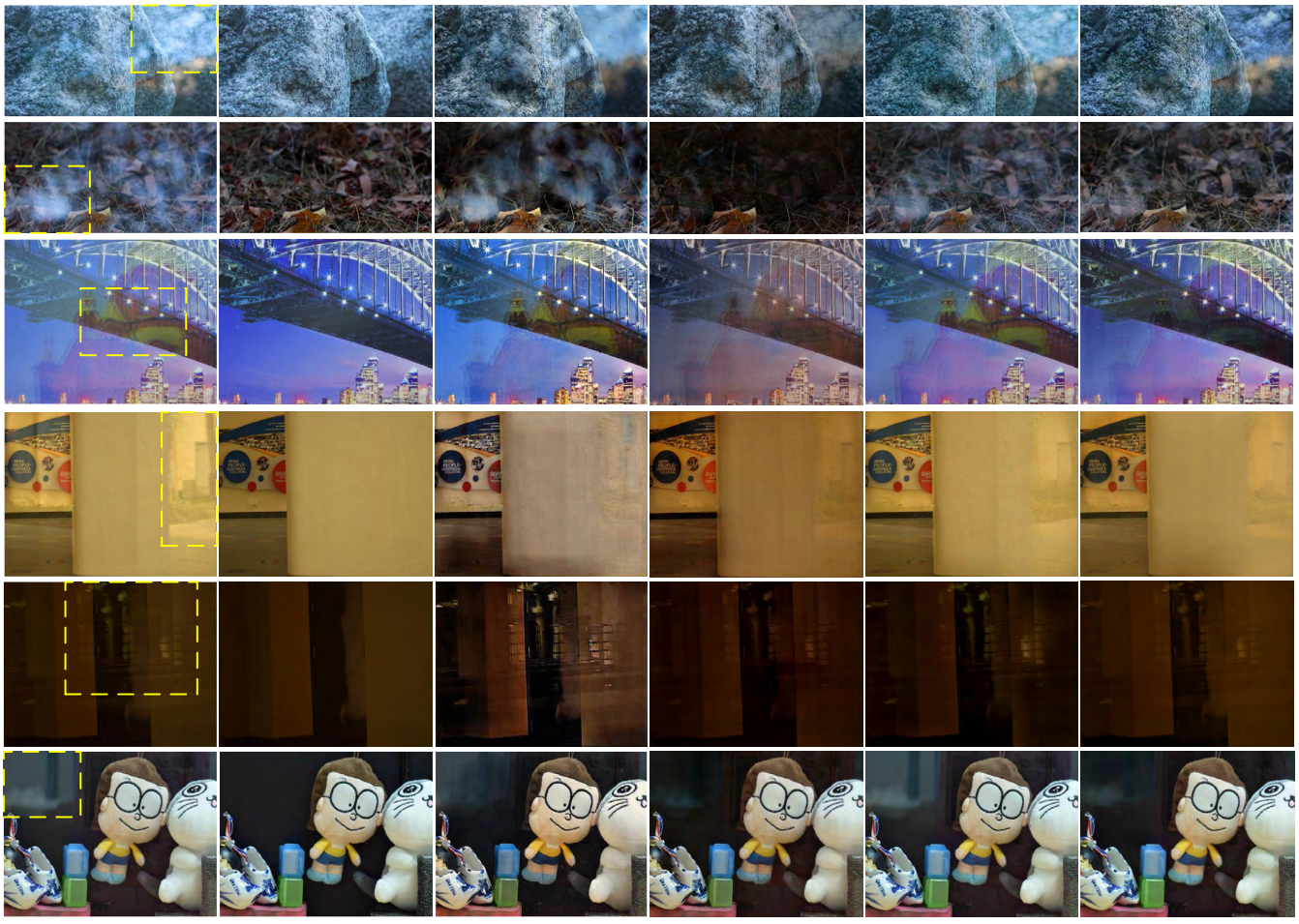


FIGURE 5. Results of different methods. (a) Input images, (b) Ground truth, (c) CEILNet results, (d) Zhang's results, (e) ERRNet results, (f) Our results.

module from the generator respectively, those three models are referred as SN-N, SP-N and CB-N.

We evaluate these models and summarize the average PSNR and SSIM of the 4 group datasets in Table 2. It can be seen that, among these 8 metrics, our algorithm achieves 5 highest scores and 3 second-best ones, which surpasses all the other methods and shows the best performance. The results of SN-N, SP-N and CB-N demonstrate the effectiveness of our guidance SNGAN and CBAM module. It should be noticed that there is still room for improvement especially on the Wild dataset, because the quality metrics of the input are higher than that of the output.

In order to further compare different algorithms through visual assessment, we present the input, ground truth, and results of several test images in Fig. 5. The first and second rows are from Berkeley dataset, others come from the

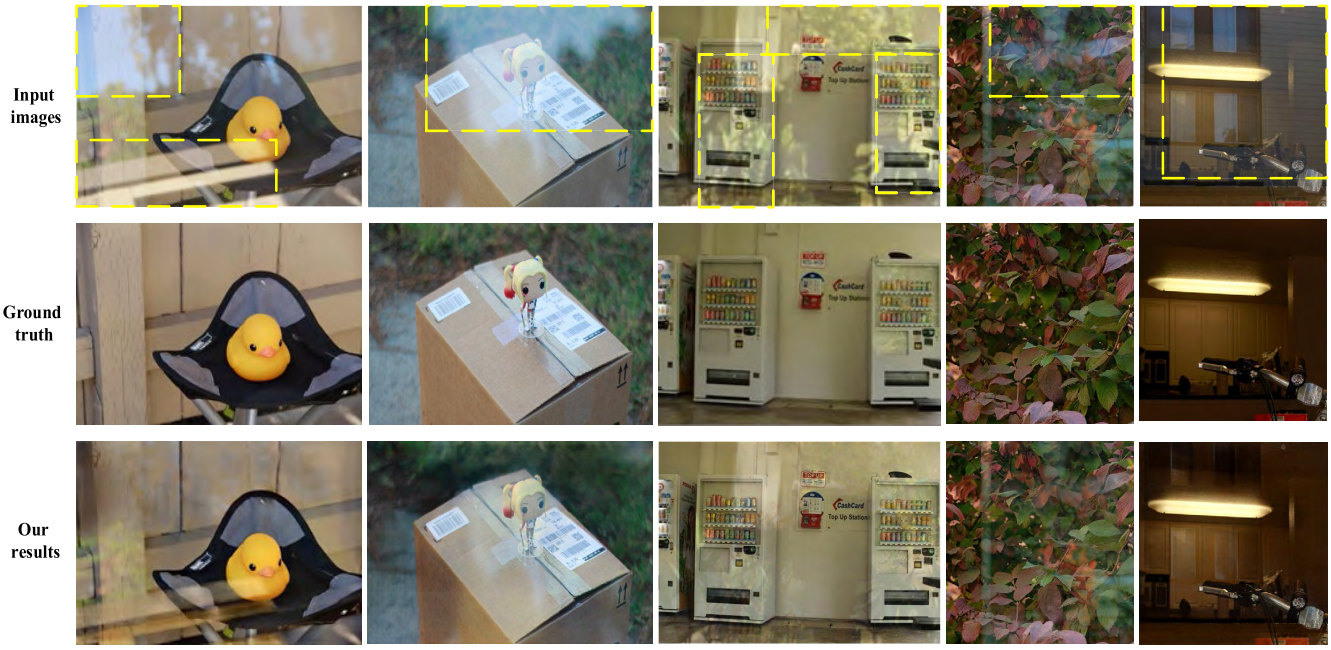


FIGURE 6. Results of the proposed method on Berkeley dataset.

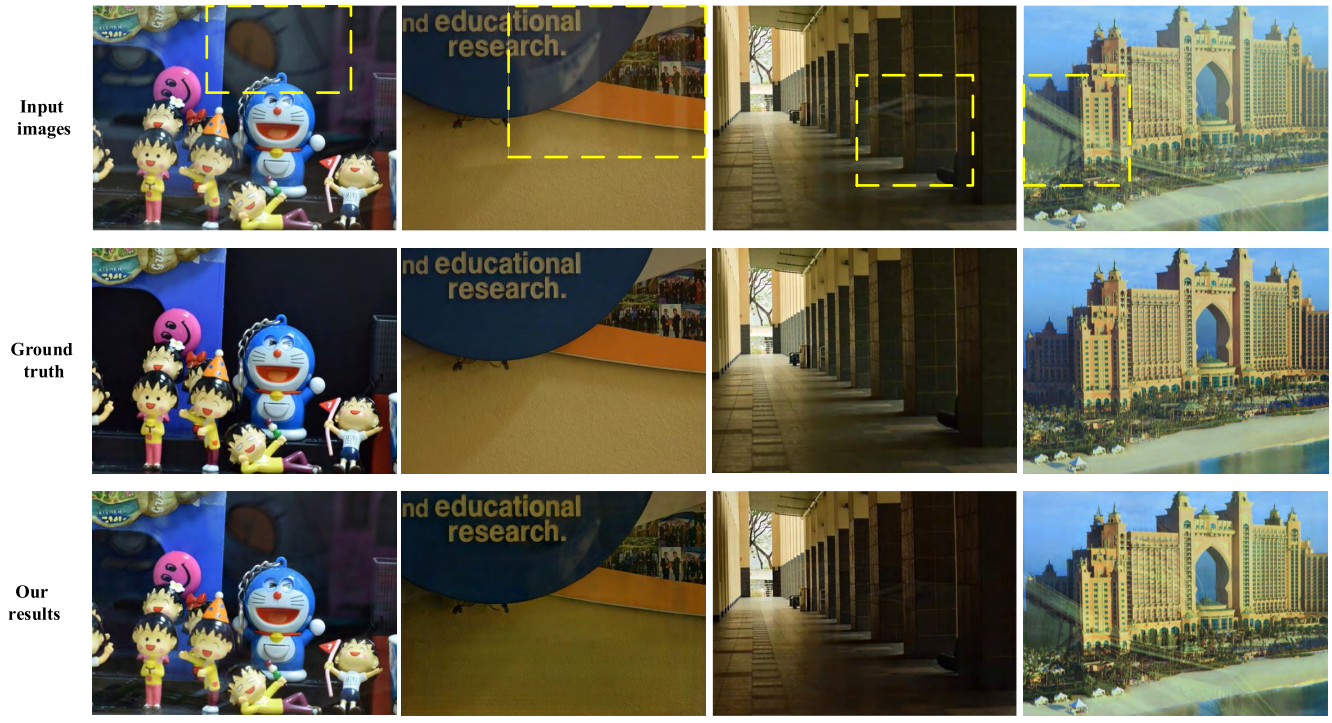


FIGURE 7. Results of the proposed method on SIR2 dataset.

SIR2 dataset. These 6 images as referred as stone, grass, bridge, lobby, door, and toy, respectively. Besides, some local patches with strong reflection signals are marked by yellow rectangles to make the evaluation more intuitive.

As can be seen in Fig. 5, the CEILNet cannot suppress the reflection layer very well, for the white reflection parts in the stone, toy, and grass images can be clearly noticed. Besides, the wall of the church overlapped on the bottom of the bridge is also very obvious, and the lobby and door images become even worse than the input images. Zhang's network removes most of the reflection signals, and the visual effect of the toy image is the best one of those models. However, some of

the results suffer from colour distortion, where the grass and door images are over dark, and the blue-sky area of the bridge image seems greyish. ERRNet performs better than CEILNet in reflection removal and doesn't involve colour distortion. Whereas the residual reflection signal still exists, especially for the stone, grass, and toy images where white reflection parts can also be distinguished by naked eyes. Additionally, the structure of the reflection layers in the bridge and lobby images are also clear, and the door image seems darker than the input. Our method has a better visual effect compared to the above methods, most of the reflection layers have been successfully removed without colour distortion. But it should be noticed that some residual artifacts can still be optimized, such as the white parts in stone and grass images. Besides, the gradient in the bridge image is still very obvious, we suspect that it is caused by the synthesizing model used in our work. The reflection is considered to be a Gaussian blurred image, which may not be rational for those inputs with sharp edges in the reflection layer, and this issue should be further addressed.

More results of the proposed network are shown in Fig. 6 (Berkeley dataset) and Fig. 7 (SIR2 dataset), from which we can see that even some residual signals can still be observed, the visual quality has been obviously improved.

In summary, the proposed network performs better than several state-of-the-art methods both in subjective and objective respects, which makes it a very promising method for reflection removal in image processing.

We also consider the running time of different networks, the average inference time of those models on different datasets are summarized in table 3. We use the official codes from the papers, the CEILNet is written in lua with torch, others are based on python and TensorFlow. The results indicate that our model runs faster than other state-of-the-art methods.

TABLE 3. Running time on benchmark datasets (unit: s).

Dataset	CEILNet	Zhang's	ERRNet	Proposed
Real	0.222	0.147	0.369	0.117
Wild	0.237	0.127	0.279	0.104
SolidObject	0.181	0.109	0.108	0.086
Postcard	0.248	0.108	0.107	0.085

V. CONCLUSION

This paper introduces an improved single image reflection removal algorithm. The key point of this method is to exploit the non-uniformity of the reflection-relevant information in both generative and discriminative networks via attention model and SN-GAN with reflection mask. Experimental results indicate that the proposed network is superior to some state-of-the-art methods in terms of PSNR, SSIM, and visual assessment. Yet there is still room for improvement to further remove the residual signal and supress the gradients of sharp reflection layer.

ACKNOWLEDGMENT

The authors would like to appreciate Professor Adams Kong for the facilities he provided. The SIR2 dataset is made available by the ROSE Lab at NTU. Portions of the research in this paper were done at the Nanyang Technological University (NTU), Singapore.

REFERENCES

[1] T. Xue, M. Rubinstein, C. Liu, W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, p. 79, 2015.

[2] Y. Li and M. S. Brown, "Exploiting reflection change for automatic reflection removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2432–2439.

[3] R. Fergus, B. Singh, A. Hertzmann, S. T. Roweis, and W. T. Freeman, "Removing camera shake from a single photograph," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 787–794, Jul. 2006.

[4] Q. Shan, J. Jia, and A. Agarwala, "High-quality motion deblurring from a single image,"*ACM Trans. Graph.*, vol. 27, no. 3, p. 73, 2008.

[5] Gong, Yuanhao, and Ivo F. Sbalzarini, "Image enhancement by gradient distribution specification," in *Proc. Asian Conf. Comput. Vis.*, 2014, pp. 47–62.

[6] A. Levin, A. Zomet, and Y. Weiss, "Separating reflections from a single image using local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun./Jul. 2004, pp. 1–8.

[7] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1647–1654, Sep. 2007.

[8] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2004, pp. 602–613.

[9] Y.-C. Chung, S.-L. Chang, J.-M. Wang, and S.-W. Chen, "Interference reflection separation from a single image," in *Proc. Workshop Appl. Comput. Vis. (WACV)*, Dec. 2009, pp. 1–6.

[10] Q. Yan, Y. Xu, X. Yang, and T. Nguyen, "Separation of weak reflection from a single superimposed image," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1173–1176, Oct. 2014.

[11] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot, "Depth of field guided reflection removal," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 21–25.

[12] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3193–3201.

[13] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3238–3247.

[14] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "CRRN: Multi-scale guided concurrent reflection removal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4777–4785.

[15] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4786–4794.

[16] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8178–8187.

[17] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3771–3779.

[18] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[19] K. Zeng, J. Yu, R. Wang, C. Li, and D. Tao, "Coupled deep autoencoder for single image super-resolution," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 27–37, Jan. 2017.

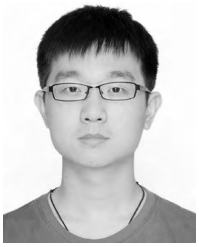
[20] Mao, Xiaojiao, Chunhua Shen, and Yu-Bin Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2802–2810.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[22] L.-C. Chen, Y. Zhu, and G. Papandreou, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[23] F. Gao, Y. Yang, J. Wang, J. Sun, E. Yang, and H. Zhou, "A deep convolutional generative adversarial networks (DCGANs)-based semi-supervised method for object recognition in synthetic aperture radar (SAR) images," *Remote Sens.*, vol. 10, no. 6, p. 846, May 2018.
[24] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, *arXiv:1701.07875*. [Online]. Available: http://arxiv.org/abs/1701.07875
[25] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
[26] A. Almahairi, N. Ballas, and T. Cooijmans, "Dynamic capacity networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2549–2558.
[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[28] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[29] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2746–2750.
[30] A. Brock, T. Lim, J. M. Ritchie, and N. Weston, "Generative and discriminative voxel modeling with convolutional neural networks," 2016, *arXiv:1608.04236*. [Online]. Available: http://arxiv.org/abs/1608.04236
[31] G. Ian, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, vol. 2, 2014, pp. 2672–2680.
[32] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.
[33] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3922–3930.
[34] R. Lai, Y. Li, J. Guan, and A. Xiong, "Multi-scale visual attention deep convolutional neural network for multi-focus image fusion," *IEEE Access*, vol. 7, pp. 114385–114399, 2019.
[35] J. Guan, R. Lai, A. Xiong, Z. Liu, and L. Gu, "Fixed pattern noise reduction for infrared images based on cascade residual attention CNN," *Neurocomputing*, vol. 377, pp. 301–313, Feb. 2020, doi: 10.1016/j.neucom.2019.10.054.
[36] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[37] X. Du, S. Yin, R. Tang, Y. Zhang, and S. Li, "Cardiac-DeepIED: Automatic pixel-level deep segmentation for cardiac bi-ventricle using improved end-to-end encoder-decoder network," *IEEE J. Transl. Eng. Health Med.*, vol. 7, 2019, Art. no. 1900110.
[38] R. Wan, B. Shi, H. Li, L.-Y. Duan, A.-H. Tan, and A. K. Chichung, "CoRRN: Cooperative reflection removal network," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 7, 2019, doi: 10.1109/TPAMI.2019.2921574.
[39] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 654–669.
[40] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: http://arxiv.org/abs/1802.05957
[41] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," 2019, *arXiv:1911.06634*. [Online]. Available: http://arxiv.org/abs/1911.06634

KUANHONG CHENG was born in Henan, China, in 1990. He received the B.S. degree from Xidian University, in 2012, where he is currently pursuing the Ph.D. degree with the School of Physics and Optoelectronics Engineering. From 2018 to 2019, he was a Visiting Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests are image processing and computer vision.

JIANGLUQI SONG received the Ph.D. degree in condensed matter physics from the University of Science and Technology of China, in 2017. He currently works as a Senior Lecturer with Xidian University. His research interests include 3D reconstruction, hyperspectral image processing, biophotonics, and targeted diagnosis.

JUAN DU was born in Shaanxi, China, in 1990. She received the B.S. degree from Xi'an Technological University, in 2013. She is currently pursuing the Ph.D. degree with the School of Physics and Optoelectronics Engineering, Xidian University. Her major research interests include image processing, hyperspectral imaging, and image super resolution based on deep learning.



SHENGHUI RONG was born in Rizhao, China, in 1989. He received the B.S. degree in electronic science and technology and the Ph.D. degree in physical electronics from Xidian University, Xi'an, China, in 2011 and 2018, respectively. In 2016, he was funded by the China Scholarship Council (CSC) to conduct his research in the directions of 3D image processing and recognition at Griffith University, Australia. He is currently a Lecturer with the School of Information Science and Engineering, Ocean University of China (OUC). His primary research interests include optoelectronics countermeasures, computer vision, and pattern recognition.

HUIXIN ZHOU (Member, IEEE) received the Ph.D. degree from Xidian University, Xi'an, China, in 2004. He is a Professor with Xidian University. His current research areas include optoelectronics imaging and real-time image processing, target detecting and tracking, and high/hyperspectral image processing. He is a Senior Member of the Photoelectronics Technology Professional Committee, Chinese Society of Astronautics, and the Chinese Optical Society, and also a member of The Optical Society of America.