Multi-Scale Frequency Separation Network for Image Deblurring

Yanni Zhang, Qiang Li, Miao Qi, Di Liu, Jun Kong, and Jianzhong Wang

Abstract—Image deblurring aims to restore the detailed texture information or structures from the blurry images, which has become an indispensable step in many computer vision tasks. Although various methods have been proposed to deal with the image deblurring problem, most of them treated the blurry image as a whole and neglected the characteristics of different image frequencies. In this paper, we present a new method called multi-scale frequency separation network (MSFS-Nets) for image deblurring. MSFS-Net introduces the frequency separation module (FSM) into an encoder-decoder network architecture to capture the low- and high-frequency information of image at multiple scales. Then, a simple cycle-consistency strategy and a sophisticated contrastive learning module (CLM) are respectively designed to retain the low-frequency information and recover the high-frequency information during deblurring. At last, the features of different scales are fused by a cross-scale feature fusion module (CSFFM). Extensive experiments on benchmark datasets show that the proposed network achieves state-of-the-art performance.

Index Terms—Image deblurring, frequency separation, contrastive learning, cross-scale feature fusion.

I. INTRODUCTION

THE blur artifact caused by motion of camera/object or out-of-focus lens will affect the image quality and severely degrade the performance of downstream computer vision tasks, such as video surveillance, object detection and face recognition. Therefore, accurate and efficient image deblurring techniques which could restore the sharp images from their blurry counterparts have attracted much attention in both academic and industrial communities.

In the early studies, most image deblurring methods focused on estimating the blurry kernel by introducing some prior information [1]. However, since the blur in an image may be induced by multiple reasons, image deblurring becomes a highly ill-posed problem and it is difficult to model the complex blur kernel by simple and linear assumptions.

With the continuous development of deep learning, some deep convolutional neural networks (CNNs) [2], [3], [4], [5], [6], [7] have been adopted as blur kernel estimator and showed satisfied deblurring performance. However, these methods always need two stages to accomplish the image deblurring task. That is, they first interpolated the blur kernel using CNN and then utilized the estimated kernel for blurry image deconvolution. Therefore, the above methods may suffer from both high computational burden and inaccurate blur kernel estimation. More recently, some other CNN-based image deblurring methods were proposed to directly learn the relationship between blurry and sharp images by an image-to-image regression manner [8], [9], [10], [11], [12], [13]. Compared with other works, the advantage of image-to-image regression methods is that they could avoid the deblurring errors induced by inadequate blur kernel estimation. Besides, the CNN has also been combined with some other techniques for image deblurring. In [11], [14], [15], Recurrent Neural Network (RNN) was considered as a deconvolution operation to model the spatially variant blur. Inspired by Generative Adversarial Network (GAN), researchers in [16], [17] employed the generative and discriminative models into the image deblurring framework, which could make the deblurred images more realistic and effectively improve the deblurring quality.

Although the aforementioned CNN based deblurring methods adopted various techniques to remove the blur from images, most of them adopted the encoder-decoder architecture to capture multi-scale image features. That is, they first leverage encoder to gradually reduce the input blurry image to low-resolution representations, and then utilize decoder to progressively recover the original resolution for deblurring. This multi-scale strategy is reasonable for image deblurring because it could treat different image features in different scales. The coarse image features can be captured by low-resolution representations, while the fine image features are more suitable to be recovered in high-resolution representations. However, the differences of image information not only exist in the resolution scale aspect but also can be reflected by different frequencies. That is, the smoothly changing areas and outline of an image are mainly described by its low-frequency (LF) component, while the rapidly changing fine details of the image are usually described by its high-frequency (HF) component. Therefore, since the existing CNN based image...
image deblurring network with selective parameter sharing. Instead of stacking multiple sub-networks, Purohit et al. [19] proposed a coarse-to-fine network architecture with three sub-networks to restore the sharp image in a coarse-to-fine manner, in which different sub-networks process images of different sizes. Inspired by MSCNN, Gao et al. [18] proposed a coarse-to-fine image deblurring network with selective parameter sharing and nested skip connections between different sub-networks. Instead of stacking multiple sub-networks, Purohit et al. [19] adopted an encoder-decoder backbone (RADN) with dense deformable module (DDM) and self-attention (SA) module to improve the deblurring performance without significantly increasing the computational cost. Cho et al. [9] presented a multi-input multi-output U-net (MIMO-U-Net) which utilizes a single U-Net (i.e., encoder-decoder with short connections) but multiple input and output images to handle the coarse-to-fine image deblurring. Chi et al. [20] utilized an encoder-decoder network to extract multi-scale image features, and then integrated the auxiliary and meta learning to enhance the deblurring performance. Chen et al. [13] also applied encoder-decoder architecture to implement multi-scale and multi-stage image restoration tasks by introducing a new normalization method. In order to achieve better deblurring effect, RNN is introduced for deblurring task. Zhang et al. [15] proposed an image deblurring approach by combining CNN with RNN, in which four RNN layers are utilized to receive different directional sequence of CNN features. Tao et al. [11] proposed a scale-recurrent network (SRN) by introducing the long-short term memory (LSTM) and ResBlock into an encoder-decoder based deblurring model. The success of GAN also promoted image deblurring research. Kupyn et al. proposed a DeblurGAN [16] to model different blur sources, in which a CNN with encoder-decoder architecture is employed as generator and a convolutional patch-based classifier is adopted as discriminator. Based on DeblurGAN, DeblurGAN-v2 [17] was proposed to incorporate a double-scale discriminator and a feature pyramid network into GAN to achieve better deblurring result.

B. Frequency Separation

An image can be decomposed into different frequency bands, and different frequency bands contain structures and textures with distinct complexities. Therefore, analyzing the image feature in frequency domain is a commonly used technique in many conventional low-level computer vision tasks. Recently, researchers have also proposed some deep learning based deblurring methods which consider the characteristics of different image frequency. [3], [4] employed CNN for blur kernel estimation in frequency domain and achieve satisfactory results. In image-to-image regression framework, Liu et al. [12] designed a two-stage method which first separates the high-frequency residual information from the blurry image and then adopt an encoder-decoder network to realize the high frequency information refinement. Zou et al. [10] utilized discrete wavelet transform to divide the dilated convolution features into four frequency bands, so that different frequency features can be refined independently. Nevertheless, the above two methods only separate the image frequency in the first or last layer of the network. Thus, they can only capture the image features of different frequencies from a specific scale and ignored the different image frequency features of multiple scales. More recently, the spatial frequency domain based method has also been integrated with imaging sensor to deal with aberration-correction problem [21].

C. Contrastive Learning

Contrastive learning [22] is a widely used self-supervised strategy and has become an effective tool to handle some
real-world problems. Motivated by the success of its application in representation learning [23], some researchers have adopted contrastive learning to model the comparative relationships between features for computer vision tasks [24], [25]. Recently, Park et al. adopted contrastive learning in an image-to-image translation network [26]. Wu et al. [27] designed a network using contrastive learning to remove the haze from hazy image. Wang et al. [28] also applied contrastive learning to obtain invariant degradation representation in image super-resolution problem. Although these methods demonstrated that contrastive learning can help to improve the performance of some low-level vision tasks, there are few works employ contrastive learning in image deblurring problem. Therefore, how to make good use of contrastive learning to facilitate the performance of image deblurring is still needed to be studied.

III. METHOD

A. Motivation

In multi-scale and hierarchical image deblurring methods, researchers have realized that the images of different scales or spatial resolutions reflect diverse features [8], [9], [11], [12]. That is, the image with large scale and high spatial resolution contains fine feature while the image with small scale and low spatial resolution captures the coarse feature. Nevertheless, previous deblurring works seldom took the frequency information of image into consideration. In this study, we observe that the difference between blurry and sharp images lies in both the scale and frequency aspects. Take the images in Fig. 1 as an example, it can be seen that the low and high-frequency parts of an image at various scales exhibit very different features. In other words, the image with original size contains more detailed texture than the images with smaller sizes. Moreover, we can also find that the discrepancy between blurry and sharp images at the same scale is mainly reflected by their high-frequency components. Specifically, the high-frequency component of a sharp image always captures more clear texture and structures than the high-frequency component of a blurry image at the same scale, while the low-frequency components of the sharp and blurry images at the same scale are very similar. This phenomenon may due to that blurring can be regarded as a process of diffusing the information encoded in sharp edges across an image, which would not dramatically alter the smoothly changing structure and outline of the image [1].

To further justify our observation, we compare the entropy obtained by all training samples with different frequencies and scales in GoPro dataset [8]. From the distributions in Fig. 2, we can find the similar observations with those in Fig. 1. That is, the Jensen–Shannon divergences between entropy distributions obtained by high-frequency components of sharp and blurry images are much larger than those obtained by low-frequency components of sharp and blurry images. Moreover, we can also see that the difference between entropy distributions of sharp and blurry images with large scale is greater than that with small image scale.

B. Overview

Motivated by the observation in subsection III-A, we propose a multi-scale frequency separation network (MSFS-Net), which makes full use of different frequency features at different scales, to achieve better deblurring performance. Figure 3 shows the overall architecture of the MSFS-Net.

As can be seen from Fig. 3, the architecture of MSFS-Net is based on an encoder-decoder structure to hierarchically extract multi-scale image features. Firstly, a blurry image is input and a $3 \times 3$ convolution is applied to get shallow features. Then, the down-sampling module and frequency separation module (FSM) are combined in the encoder stage to progressively extract the low and high-frequency features of image at different scales. The down-sampling module consists of $3 \times 3$ convolution with step 2 and LeakyRelu, and FSM is proposed to decompose the down-sampled features into different frequency. After the encoder stage, we can get the latent feature of the input blurry image. In order to refine the latent feature, multiple RCABs [29] are further adopted to process the feature and improve the model capacity. Next, we use up-sampling module to achieve scale restoration of features in the decoder stage. The up-sampling module consists of RCAB

![Fig. 1. Detailed information of images from GoPro dataset with different frequencies and scales. (a) blurry image, (b) sharp image. From top to down are the original image, the high and low-frequency components of an image patch (marked as red box) with original, 1/2 and 1/4 scales, respectively. In this figure, the low-frequency component of each image is obtained by subtracting the low-frequency component from the original image.](image1.png)

![Fig. 2. The distributions of entropy obtained by training samples in GoPro dataset. Top: From left to right are the distributions of entropy obtained by high-frequency components of sharp ($H_{\text{Sharp}}$) and blurry ($H_{\text{Blur}}$) images at original, 1/2 and 1/4 scales. Down: From left to right are the distributions of entropy obtained by low-frequency components of sharp ($L_{\text{Sharp}}$) and blurry ($L_{\text{Blur}}$) images at original, 1/2 and 1/4 scales.](image2.png)
Fig. 3. The architecture of the proposed MSFS-Net. The encoder and decoder are indicated by red and blue boxes, respectively. Note that the two encoders are identical and share the same parameters.

with pixel-shuffle [30] and FSM is also adopted to decompose the restored features at each scale. Since the decoder stage requires delicately use of fine-grained details to reconstruct features, the cross-scale feature fusion module (CSFFM) is applied to connect features at different scales of encoder and decoder stages so that different context information can be passed to each other and well preserved. In order to minimize the loss of information and make the network converge rapidly, we fuse the original input image with the features after the last $3 \times 3$ convolution of decoder by an element-wise summation. Last but most important, in order to take full advantage of low and high-frequency information, we reuse the encoder of network to obtain different frequency features of sharp image at different scales, and two distinct strategies are carried out to constrain the low and high-frequency features in the intermedia layers of our network. On the one hand, since low-frequency features of the blurry and sharp images at the same scale are similar, a simple cycle-consistency criterion is utilized to ensure that the low-frequency features of the sharp and input blurry images are not far away from each other. On the other hand, we propose a contrastive learning module (CLM) to regularize the high frequency features in the decoder stage. In CLM, we regard the high frequency features of encoder, decoder and generated sharp image at the same scale as negative, anchor and positive, respectively. Through the two opposite forces in CLM which pull the anchor closer to positive point and push the anchor farther away from negative point in the feature space, the interference of high frequency features in blurry image can be effectively removed. Here, it should be noted that we utilize the low and high-frequency components of sharp image to constrain the intermedia features of different stages in the backbone network (i.e., low-frequency for encoder constraint and high-frequency for decoder constraint). This is because that the encoder is mainly used to extract context and outline information of the blurry image while the detailed information of sharp image is mostly generated by the decoder. Moreover, the cycle-consistency and CLM introduce multiple closed-loop structure in our network, which is helpful to reduce the solution space of our model [31].

C. Frequency Separation Module

The natural image can be divided into low and high-frequency components, and the output of a convolution layer can also be decomposed into features with different spatial frequencies. In Octave Convolution (OctConv) [32] and some other studies [33], researchers have shown that the flat and smoothly changing areas with spatial redundancy (i.e., low-frequency information) of an image can be easily captured in low resolution feature maps, while the high resolution feature maps are more appropriate to characterize the rapidly changing details and fine boundaries (i.e., high-frequency information) in image. Thus, OctConv is used as the basic block of our frequency separation module (FSM). The structure of OctConv is shown in Fig. 4(a). Suppose $X \in \mathbb{R}^{c_{in} \times h \times w}$ is the input feature in which $h$ and $w$ denote the spatial dimensions and $c_{in}$ is the number of channels. OctConv first decomposes $X$ into two parts, one is $X^H \in \mathbb{R}^{(1-\alpha_{in})c_{in} \times 0.5h \times 0.5w}$ to capture high-frequency information, and the other is $X^L \in \mathbb{R}^{\alpha_{in}c_{in} \times 0.5h \times 0.5w}$ for low-frequency information extraction. The parameter $\alpha_{in}$ adjusts the number of channels in low and high-frequency. Then, the low and high-frequency features are processed by convolution and information interaction between two frequen-
ties will be carried out through pooling and up-sampling operations. The process of OctConv can be expressed by the following equations:

\[
Y^H = f(X^H; W^H \rightarrow H) + \text{upsample}(f(X^L; W^L \rightarrow H), 2)
\]

(1)

\[
Y^L = f(X^L; W^L \rightarrow L) + f(\text{pool}(X^H, 2); W^H \rightarrow L)
\]

(2)

where \(f(X; W)\) represents the convolution with kernel \(W\), and the convolution kernel \(W\) is divided into \(W^H\) and \(W^L\) to convolve with \(X^H\) and \(X^L\) respectively. \(W^H\) can be further divided into \(W^H \rightarrow H\) and \(W^H \rightarrow L\) for intra- and inter-frequency processing. Similarly, \(W^L\) can also be divided into \(W^L \rightarrow L\) and \(W^H \rightarrow L\). This process can realize the communication of low or high-frequency information. To deal with the mismatch between spatial scales of \(X^H\) and \(X^L\), \(\text{pool}(X, 2)\) and \(\text{upsample}(X, 2)\) are used. \(\text{pool}(X, 2)\) represents average pooling with kernel size 2 × 2 and stride 2, and \(\text{upsample}(X, 2)\) is an up-sampling operation by a factor of 2. Through the above operations, the output high-frequency feature \(Y^H \in \mathbb{R}^{(1-\alpha_{in}) \times \alpha_{out} \times h \times w}\) and low-frequency feature \(Y^L \in \mathbb{R}^{\alpha_{in} \times \alpha_{out} \times 0.5h \times 0.5w}\) can be obtained. The parameter \(\alpha_{out}\) also adjusts the output channel \(c_{out}\).

Based on OctConv, the proposed FSM is shown in Fig. 4(b). First, a \(1 \times 1\) OctConv (\(\alpha_{in} = 0, \alpha_{out} = 0.5\)) is utilized to divide the input feature into low and high-frequency parts. Then a \(3 \times 3\) OctConv (\(\alpha_{in} = 0.5\) and \(\alpha_{out} = 0.5\)) is applied to accomplish the information exchange and updating. As a result, the low-frequency feature \(Y^L\) and high-frequency feature \(Y^H\) (shown in the yellow box in Fig. 4(b)) of the input feature can be obtained. Next, a \(1 \times 1\) OctConv (\(\alpha_{in} = 0.5, \alpha_{out} = 0\)) is used to fuse the low and high frequency features into a whole for the subsequent down-sampling or up-sampling operation. At last, a residual connection is utilized to integrate the input feature with the output of the last OctConv, so that important information is not lost in FSM.

D. Cross-Scale Feature Fusion Module

From the analysis and observation in previous sections, we know that the image features with different scales exhibit different characteristics. The image feature with large scale contains fine structures such as clear edges and textures. However, with the down-sampling of feature scale, the fine structures will gradually degenerate and only the coarse structures (such as the rough contours) in the image are left. Thus, the discrepancy between the blur and sharp images with large scale is bigger than small scale (see Fig. 2).

In our study, different from other encoder-decoder based networks [8], [15], [18] which only fuse image features with the same scale in encoder and decoder stages, a cross-scale feature fusion module (CSFFM) is proposed to achieve the information fusion of features with different scales in different stages (encoder/decoder). The CSFFM is based on the idea of adaptive mix-up operation [34] and its specific process is shown in Fig. 5. The process is shown in the following equations:

\[
f_{\text{de}_{1}} = \text{up}(\left[ \text{Sigmoid}(\theta) \ast f_{\text{en}_{1}} \right], f_{\text{en}_{1}})
\]

(3)

\[
f_{\text{de}_{1}} = \text{up}(\left[ \text{Sigmoid}(\gamma) \ast f_{\text{en}_{1}} \right], f_{\text{en}_{1}})
\]

(4)

where, \(f_{\text{en}_{1}}\) and \(f_{\text{de}_{1}}\) represent the features of \(\frac{1}{4}\) scale in encoder and decoder stages (the value of \(i\) is 1, 2 and 4 in our study), \(\theta\) and \(\gamma\) are the parameters optimizable by network, \(\text{up}\) represents the up-sampling operation and \([\ ]\) denotes the concatenation. CSFFM not only connects the features of encoder and decoder, but also realize the fusion of features at different scales. At the same time, in order to retain the important features of input image, we also concatenate the fused feature with the feature of encoder.

E. Contrastive Learning Module

The main idea of contrastive learning is to pull the positive paired samples together while push negative paired samples far apart in a feature space. In our study, a contrastive learning module (CLM) is proposed to regularize the high-frequency features in decoder stage to get better restored images. According to Fig. 3, we can see that each CLM leverages three different high-frequency features at the same scale to construct the positive and negative pairs for contrast. Here, we take the high-frequency features obtained by the encoder stage, sharp image and decoder stage as negative samples, positive samples, and anchors, respectively. The reasons for this design are two-fold. First of all, the high-frequency features in each scale of encoder stage are mainly captured from the blurry image, so the information contained in them are unclear. However, since the contents in blurry and sharp images are the same, the high-frequency features of blurry image is more difficult to be distinguished from those of corresponding sharp image than other images with different content. As a result, the high-frequency features of a blurry image can be regarded as hard negative samples to promote the performance of contrastive learning. Secondly, since the sharp image contains clear image details and edges, so the high-frequency features of it can be considered as guidance for the intermedia features in the decoder stage. Through the CLMs at multiple scales and
the loss function associated with them, the adverse information in the high-frequency features of blurry image can be effectively suppressed.

F. Loss Functions

1) Multi-Scale Consistent Loss for Low-Frequency Features: Since the low frequency features often reflect the outline and rough contour information of an image. The difference between the low-frequency features of blurry and sharp images is not obvious. Inspired by cycle-consistency, we minimize the \( L_1 \) distance between the low-frequency features in encoder stage and those generated by the sharp image, so that the low-frequency features can be maintained during deblurring. Therefore, the multi-scale loss for low-frequency features can be expressed by:

\[
L_{\text{low}} = \sum_{k=0}^{2} ||f_{k}\text{en}_{\text{low}} - f_{k}\text{low}||_1
\]

where \( k \) represents scale level, \( f_{k}\text{en}_{\text{low}} \) represents the low-frequency feature of \( \frac{1}{2^k} \) scale in the encoder stage, \( f_{k}\text{low} \) represents the low-frequency feature of \( \frac{1}{2^k} \) scale got by sharp image. We adopt \( L_1 \) distance in Eq. (5) because the \( L_1 \)-norm based loss function is more suitable to measure the small difference between two similar samples and possesses good convergence property [35]. Furthermore, we should point out that since the input blurry image and sharp image are both decomposed into multi-scale low and high-frequency components by the same encoder in our model, the consistent loss of low-frequency features in Eq. (5) will lead the diversity of blurry and sharp images to be mainly reflected by their high-frequency components at each scale, which could also facilitate our contrastive learning module for high-frequency features.

2) Multi-Scale Contrastive Loss for High-Frequency Features: As mentioned in contrastive learning module, we introduce the contrastive learning module (CLM) into our MSFS-Net for better image deblurring. Therefore, considering all CLMs at different scales, the multi-scale contrastive loss \( (L_{\text{high}}) \) for regularizing the high-frequency features in the network can be expressed by the following equation:

\[
L_{\text{high}} = \sum_{k=0}^{2} \frac{L_1(f_{k}\text{anchor}, f_{k}\text{positive})}{L_1(f_{k}\text{anchor}, f_{k}\text{negative})}
\]

where \( k \) represents the scale level of feature, \( f_{k}\text{anchor} \) represents the high-frequency feature of \( \frac{1}{2^k} \) scale in the decoder stage, \( f_{k}\text{positive} \) represents the high-frequency feature of \( \frac{1}{2^k} \) scale got by sharp image, \( f_{k}\text{negative} \) represents the high-frequency feature of \( \frac{1}{2^k} \) scale in the encoder stage, \( L_1 \) represents the \( L_1 \)-distance. In Eq. (6), the numerator and denominator are utilized to pull the \( f_{k}\text{anchor} \) and \( f_{k}\text{positive} \) together and push the \( f_{k}\text{anchor} \) and \( f_{k}\text{negative} \) apart, respectively.

3) The Final Loss of MSFS-Net: At last, the total loss function used to train our MSFS-Net can be defined as:

\[
L_{\text{total}} = \lambda_1 L_{\text{high}} + \lambda_2 L_{\text{low}} + ||I - G||_1
\]

where \( \lambda_1 = \lambda_2 \) are set as 0.005 by experiment, \( I \) represents the sharp image obtained by our network and \( G \) is the ground-truth, \( L_1 \) norm is applied to minimize the loss between the recovered image and ground-truth.

G. Comparison With Other Methods

To highlight the novelty of the proposed model, we compare MSFS-Net with some related methods. First, though the frequency separation has been adopted in some works to deal with image restoration problem such as super-resolution [36], [37] and deraining [38], they only decomposed different frequency information from image features at a specific size. Thus, the characteristics of image features at different scales are neglected in them. For the methods which considered the frequency information of images in deblurring task [10], [12], they either only focused on high-frequency features or indiscriminatingly treated different frequency features of an image with the same strategy. Hence, they are still different from our proposed method. Besides, unlike some of the aforementioned methods that embed discrete cosine transform (DCT), discrete wavelet transform (DWT) and their inverse operations into the network for frequency analysis, the pure convolution based network architecture of our MSFS-Net can avoid information interchanges between the frequency and spatial domains, which make the information propagate more smoothly. Second, contrastive learning has also been employed in some image-to-image regression tasks [26], [27]. However, these methods merely leveraged the contrastive learning to regularize the final output rather than the intermedia layers of the network. The different frequencies of multi-scale image features are ignored in them. The last technique related to our work is the perceptual loss [39] which also utilizes a multi-layer network to extract the features of network’s output. Nevertheless, the differences between perceptual loss and our work are still apparent. The aim of perceptual loss is to measure the visual difference between the network’s output and the ground-truth by features extracted from a pre-trained deep neural network (i.e., VGG [40]). Hence, it cannot be adopted to constrain the features obtained by intermedia layers of backbone network. Furthermore, perceptual loss also overlooks the different frequency information of the image.

IV. EXPERIMENTS

A. Dataset and Implementation Details

To comprehensively validate the proposed MSFS-Net for various image deblurring problems, we first use GoPro [8],
ZHANG et al.: MULTI-SCALE FREQUENCY SEPARATION NETWORK FOR IMAGE DEBLURRING

Fig. 6. Visual comparison of the deblurring results on GoPro dataset.

HIDE [41] and RealBlur [42] datasets to test the performance of our model for motion deblurring task. Then, DPDD [43] dataset is employed to evaluate the effectiveness of our model for defocus deblurring task. GoPro dataset contains 3214 pairs of blurry and sharp images, in which the training and test sets consist of 2103 and 1111 pairs, respectively. HIDE dataset consists of 8422 pairs of blurry and sharp images and these images are carefully selected from 31 high-fps videos. RealBlur dataset consists of two subsets: RealBlur-J and RealBlur-R. DPDD dataset provides 500 defocused and all-in-focus image pairs. The training, validation, and testing sets in this dataset consist of 350, 74, and 76 image pairs, respectively. Following the experiments in other works [44], [45], [46], we train and test our MSFS-Net by the training and test sets of DPDD without using the dual-pixel images. For implementation details, the AdamW optimizer with parameter setting as $\beta_1 = 0.9$, $\beta_2 = 0.9$, $\epsilon = 1e-8$ is used to optimize our network. The epochs and batch size are set as 3000 and 4 respectively. The initial learning rate is set as $1e-4$ and decreased by the factor of 0.5 at every 500 epochs. Note that the CLM, consistency and encoder of sharp image in our MSFS-Net are removed after training, and only the backbone (i.e. encoder-decoder structure with CSFFM) is left for inference.

B. Quantitative and Qualitative Evaluation

1) Motion Deblurring: To demonstrate the effectiveness of the proposed MSFS-Net, we compare the performance of our method with some state-of-the-art algorithms on GoPro dataset. The quantitative comparison results are listed in Table I, and some visual comparisons are shown in Fig. 6. In our experiment, the PSNR and SSIM results of all comparison methods are directly quoted from their corresponding literatures. Recently, Chu et al. [47] have shown that Test-time Local Converter (TLC) can effectively reduce the inconsistency of train-test information distributions and improve the performance of image restoration without any model fine-tuning. Thus, we also combine TLC with our proposed MSFS-Net (denoted as MSFS-Net-Local) to compare its deblurring performance with some improved versions of other methods.

From Table I, we can see that the deblurring performance of our method is superior to other state-of-the-art methods. The advantage of MSFS-Net can be attributed to the following reasons. First, although some comparison methods [9], [11] utilize the encoder-decoder network structure to extract multi-scale features of the image for deblurring, they only adopt a simple skip connection mechanism to concatenate the features with the same scale in encoder and decoder. Nevertheless, thanks to the CSFFM in our model, the features of different stages (i.e., encoder and decoder) with different scales can be better fused. Therefore, the proposed MSFS-Net performs better than them. Second, the methods in [8], [16] integrate the image features with different scales by some sophisticated network structure and modules. However, they treat the image as a whole and neglect the characteristics of different image frequencies. Thus, their performance is inferior to our model which makes full use of the low- and high-frequency information separated by FSM. Third, the method in [12] considers the frequency information in image deblurring problem. But it only focuses the high-frequency image features. Although SDWNNet [10] utilizes the wavelet transformation for image frequency separation, it indistinguishably processes the low- and high-frequency image features using the same network. Hence, the deblurring results obtained by the above two methods are still not optimal. Besides, the methods in [10] and [12] only capture image frequency information at the
first or last layer of the network and ignore the multi-scale frequency features of the image, which may also decrease their performance. At last, the proposed MSFS-Net can capture different frequency features of the image at multiple scales and designs different strategies (i.e., contrastive learning and consistent loss) to handle them separately. Thus, it achieves the best deblurring result. From Table I, we can also see that TLC promotes the performance of our MSFS-Net and MSFS-Net-Local outperforms the improved versions of some other methods.

Through the visual comparison in Fig. 6, we can see that our MSFS-Net outperforms other methods. For example, by comparing the deblurred images contain in the first and second lines of Fig. 6, it can be found that our model can restore more detailed structures in the face and clearer contours of the letters. Moreover, from the deblurred results in the third and fourth lines of Fig. 6, we can see that our model not only gets clearer contour of car and billboard, but also achieves better restoration for the image deformation and artifacts caused by severe object motion or camera shake.

In Fig. 7, we show the entropy distributions of low- and high-frequency components obtained by sharp images and deblurred images of our MSFS-Net. Through comparing the results with those in Fig. 2, we can see that our proposed method can effectively narrow the gap between the sharp and blurry images. That is, the Jenson-Shannon divergencies between entropy distributions of different frequencies at each scale are much smaller than those in Fig. 2.

Following some other works [8], [10], [11], we also evaluate our GoPro-trained MSFS-Net on HIDE dataset and compare its performance with other approaches. From the experimental results in Table I and visual comparisons in Fig. 8, we can clearly see the advantage of the proposed method. That is, our MSFS-Net can restore more detailed information from the blurry images and obtains better PSNR and SSIM results than other methods. According to the analysis in previous subsection, this improvement is due to the multi-scale frequency separation, contrasting learning and cross-scale feature fusion modules proposed in our method.

Besides, in order to further test the generalization ability of the proposed MSFS-Net to real blurry images, we test the performance of our method on RealBlur dataset. The advantage of our MSFS-Net for handling real blurry images is justified in Table II. Here, we should note that TLC cannot improve the performance of our GoPro trained model when it is directly applied on RealBlur dataset. This may due to the blurry images in RealBlur are captured in real scenario rather than in controlled laboratory settings.

| Methods                     | GoPro  | HIDE       |
|-----------------------------|--------|------------|
|                            | PSNR   | SSIM       | PSNR       | SSIM       |
| DeblurGAN [8]               | 28.70  | 0.858      | 24.51      | 0.871      |
| MSCNN [8]                   | 29.08  | 0.914      | 25.73      | 0.874      |
| Zhang et al. [8]            | 29.19  | 0.911      | -          | -          |
| DeblurGAN-v2 [8]            | 29.55  | 0.934      | 26.61      | 0.875      |
| Yuan et al. [8]             | 29.81  | 0.937      | -          | -          |
| DMPH [14]                   | 30.21  | 0.935      | 29.09      | 0.924      |
| SRN [11]                    | 30.26  | 0.934      | 28.36      | 0.915      |
| Lin et al. [12]             | 30.31  | 0.920      | -          | -          |
| Gao et al. [18]             | 30.92  | 0.942      | 29.11      | 0.913      |
| EADNet [49]                 | 31.02  | 0.912      | -          | -          |
| DBGAN [50]                  | 31.10  | 0.942      | 28.94      | 0.915      |
| MR-BNN [51]                 | 31.15  | 0.945      | 29.15      | 0.918      |
| SDWNet [10]                 | 31.26  | 0.966      | 28.99      | 0.957      |
| PDAGNet [52]                | 31.58  | 0.919      | -          | -          |
| Whang et al. [53]           | 31.66  | 0.948      | 29.77      | 0.922      |
| BGD [54]                    | 31.71  | 0.944      | -          | -          |
| MIMO-UNet [9]               | 31.73  | 0.951      | -          | -          |
| HLFNet [35]                 | 31.75  | 0.953      | -          | -          |
| RADN [19]                   | 31.76  | 0.953      | -          | -          |
| Jiang et al. [56]           | 31.79  | 0.949      | -          | -          |
| Sun et al. [57]             | 31.85  | 0.948      | 29.98      | 0.930      |
| Mao et al. [38]             | 31.92  | 0.953      | -          | -          |
| VBDeblur [59]               | 32.03  | 0.953      | 30.29      | 0.931      |
| SPAIR [60]                  | 32.06  | 0.953      | 30.04      | 0.945      |
| ASPDC [61]                  | 32.09  | 0.959      | 30.04      | 0.945      |
| MSA [62]                    | 32.14  | 0.956      | 29.40      | 0.926      |
| Cui et al. [20]             | 32.50  | 0.958      | 30.55      | 0.935      |
| MPPR [63]                   | 32.66  | 0.959      | 30.96      | 0.939      |
| Kit [64]                    | 32.70  | 0.959      | 30.98      | 0.942      |
| HNet [13]                   | 32.71  | 0.959      | 30.33      | 0.932      |
| MSFS-Net                    | 32.73  | 0.959      | 31.05      | 0.941      |
| MIMO-UNet++ [21]            | 32.58  | 0.959      | -          | -          |
| HND-Local [47]              | 33.08  | 0.962      | 30.66      | 0.936      |
| Wang et al. [53]            | 33.23  | 0.963      | 30.07      | 0.928      |
| MPPR-Local [22]             | 33.31  | 0.964      | 31.19      | 0.942      |
| MSFS-Net-Local              | 33.45  | 0.964      | 31.30      | 0.943      |
than synthesized from video. Thus, their characteristics are different from the training samples in GoPro. However, once the training of our model is conducted on RealBlur, MSFS-Net-Local outperforms the original MSFS-Net and some other methods.

2) Defocus Deblurring: To evaluate our MSFS-Net for defocus deblurring task, we compare its performance with other methods on DPDD dataset. From the quantitative comparison results listed in Table III and the visual comparison result shown in Fig. 9, we can find similar observations as the experiments for motion deblurring. On the one hand, thanks to the FSM, CSFFM, contrastive learning and consistent loss, the PSNR and SSIM achieved by MSFS-Net are better than other methods. On the other hand, our MSFS-Net can get more detailed information than other methods for some severely defocused areas in Fig. 9. Specifically, the shape of iron railing behind steel net (the first line in Fig. 9) and the texture of bark (the second line in Fig. 9) obtained by MSFS-Net are clearer than other methods.

3) Ablation Study and Analysis: In this section, we conduct several experiments to evaluate the effectiveness of the components proposed in our MSFS-Net. First, we directly replace FSM with the traditional convolution operation in our network, which means we don’t decompose the image into different frequency features, but just extract features from the whole image. As a result, the effectiveness of CLM and Consistency constraint without frequency separation can be tested. Second, we remove CSFFM from the network to ignore the fusion of information from different scales and layers. Then, we abandon CLM in our model, which means we don’t use the idea of contrastive learning to supervise the high frequency features in the intermedia layers. Finally, the Consistency constraint for low-frequency features is discarded.

Through the above settings, we can get six new network structures (MSFS-Net without FSM but with CLM or Consistency, without CSFFM, without CLM and without Consistency). In order to fairly compare their performance, we use the same parameter settings to train these networks. The results of ablation experiment on GoPro dataset are shown in Table IV. Besides, the number of parameters and average training time per epoch of each model are also provided. From this table, we can find that FSM, CSFFM, CLM and Consistency are all essential to our MSFS-Net. First, although removing FSM can reduce the number of parameters and training time, neither CLM nor Consistency constraint can achieve satisfied deblurring result without frequency separation. On the one hand, the energy of an image is mostly concentrated in its low-frequency components. Thus, if we neglect different frequency information and treat the image features as a whole,
the contrastive loss in Eq. (6) will be dominated by the very similar low-frequency components of blurry and sharp images. As a result, the difference between high-frequency components of blurry and sharp images, which is very crucial for image deblurring, will be overlooked. On the other hand, imposing the Consistency constraint on the unseparated frequency features will also prevent our model from restoring the high-frequency details during the deblurring. Second, the removal of CSFFM also deteriorates the performance of our model. This is due to CSFFM can make connections between features at different stages and scales, which results in a better information fusion. Finally, we can see that the absence of Consistency constraint or CLM for separated frequency features also has an adverse impact on the deblurring result of the proposed network. This justifies that constraining the low- and high-frequency features in the intermediate layers with contrastive learning and Consistency criterion are both important for improving the deblurring performance. Some visual comparison results of different ablation settings can be seen in Fig. 10.

Besides, we evaluate the sensitivity of the proposed model to the number of RCABs. As shown in Fig. 3, we employ RCABs in our backbone network for improving the feature extraction capacity. Thus, the number of them inevitably influence the deblurring performance of our MSFS-Net. In this experiment, we set the number of RCABs in bottleneck of backbone ($K$ in Fig. 11) from 2 to 10 with a step size of 2. From the results in Fig. 11, it can be seen that small number of RCABs cannot accurately recover the sharp image. For more RCABs such as 8 and 10, although they could achieve higher RSNR, the performance improvements are not apparent and at the cost of much more training epochs. As a result, $K = 6$ is a good choice for our model since it achieves better performance with less epochs.

At last, we justify the deployment of contrastive learning and cycle-consistency for different frequency features in our MSFS-Net. In Table V, Only CLM denotes we adopt contrastive learning strategy for both low and high-frequency features. Similarly, Only Consistency denotes we adopt cycle-consistency strategy to process both low and high-frequency features. CLM+Consistency means the contrastive learning and cycle-consistency are utilized to constrain high-frequency and low-frequency features, respectively. From the experimental results, it can seen CLM+Consistency get the best performance. We are not surprised that the result of Only Consistency is very bad because the cycle-consistency loss cannot restore the high-frequency information during deblurring. For Only CLM, although it achieves a comparable performance.
performance, the average training time per epoch of it is longer than CLM+Consistency. This is due to the loss for contrastive learning in Eq. (6) needs more computation than cycle-consistency loss in Eq. (5). Thus, it is reasonable to deploy different strategy for different frequency features in our MSFS-Net.

V. CONCLUSION

In this work, we propose a multi-scale frequency separation network for image deblurring (MSFS-Net). In order to make the network make full use of the low- and high-frequency information of features, we propose frequency separation module (FSM) to further separate features. At the same time, to make the features of different scales communicate with each other without losing information, cross-scale feature fusion module (CSFMM) is proposed to realize feature connection. In addition, the cycle-consistency and contrastive learning strategies are designed by analyzing the different characteristics of low- and high-frequency features between blurry and sharp images. Our proposed contrastive learning module (CLM) and consistency constraint can better restore the high-frequency features and retain the low-frequency features during the process of deblurring, respectively. Experiments on four datasets show that MSFS-Net achieves good results in image deblurring task.

In the future, we will apply our MSFS-Net to other image restoration tasks (such as deraining, dehazing and denoising) to further test its generalization ability. Besides, how to incorporate some novel backbone networks (such as Transformer) into our model to improve its feature extraction power is another interesting topic for our feature study.

REFERENCES

[1] J. Koh, J. Lee, and S. Yoon, “Single-image deblurring with neural networks: A comparative survey,” Comput. Vis. Image Understand., vol. 203, Feb. 2021, Art. no. 103134.

[2] J. Sun, W. Cao, Z. Xu, and J. Ponce, “Learning a convolutional neural network for non-uniform motion blur removal,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 769–777.

[3] A. Chakrabarti, “A neural approach to blind motion deblurring,” in Proc. Eur. Conf. Comput. Vis. Springer, 2016, pp. 221–235.

[4] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf, “Learning to deblur,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 7, pp. 1439–1451, 2016.

[5] Y.-Q. Liu, X. Du, H.-L. Shen, and S.-J. Chen, “Estimating generalized Gaussian blur kernels for out-of-focus image deblurring,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 3, pp. 829–843, Apr. 2021.

[6] F. Wen, R. Ying, Y. Liu, P. Liu, and T.-K. Truong, “A simple local minimal intensity prior and an improved algorithm for blind image deblurring,” IEEE Trans. Circuits Syst. Video Technol., vol. 31, no. 8, pp. 2923–2937, Aug. 2021.

[7] Y. Bai, H. Jia, M. Jiang, X. Liu, X. Xie, and W. Gao, “Single-image blind deblurring using multi-scale latent structure prior,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 7, pp. 2033–2045, Jul. 2020.

[8] S. Nah, T. H. Kim, and K. M. Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3883–3891.

[9] S. J. Cho, S. W. Ji, J. P. Hong, S. W. Jung, and S. J. Ko, “Rethinking coarse-to-fine approach in single image deblurring,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2021, pp. 4641–4650.

[10] W. Zou, M. Jiang, Y. Zhang, L. Chen, Z. Lu, and Y. Wu, “SDWNet: A straight dilated network with wavelet transformation for image deblurring,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2021, pp. 1995–1994.

[11] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, “Scale-recurrent network for deep image deblurring,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8174–8182.

[12] K.-H. Liu, C.-H. Yeh, J.-W. Chung, and C.-Y. Chang, “A motion deblurring method based on multi-scale high frequency residual image learning,” IEEE Access, vol. 8, pp. 66025–66036, 2020.

[13] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, “HiNet: Half instance normalization network for image restoration,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 182–192.

[14] H. Zhang, Y. Dai, H. Li, and P. Koniusz, “Deep stacked hierarchical multi-patch network for image deblurring,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5978–5986.

[15] J. Zhang, J. Pan, J. Ren, Y. Song, and M. H. Yang, “Dynamic scene deblurring using spatially variant recurrent neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 2521–2529.

[16] O. Kupyn, V. Budzan, M. Mykhailiych, D. Mishkin, and J. Matas, “DeblurGAN: Blind motion deblurring using conditional adversarial networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 8183–8192.

[17] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, “DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8878–8887.

[18] H. Gao, X. Tao, X. Shen, and J. Jia, “Dynamic scene deblurring with parameter selective shifting and nested skip connection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3848–3856.

[19] K. Purohit and A. Jagapalan, “Region-adaptive dense network for efficient motion deblurring,” in Proc. Conf. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (ICCV), Oct. 2021, pp. 11882–11889.

[20] Z. Chi, Y. Wang, Y. Yu, and J. Tang, “Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 9137–9146.

[21] J. Wu et al., “An integrated imaging sensor for aberration-corrected 3D photography,” Nature, vol. 612, pp. 62–71, Oct. 2022.

[22] R. Hadsell, S. Chopra, and Y. LeCun, “Dimensionality reduction by learning an invariant mapping,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2006, pp. 1735–1742.

[23] N. Komodakis and S. Gidaris, “Unsupervised representation learning by predicting image rotations,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2018, pp. 1–16.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 1597–1607.
W. Shi, J. Caballero, F. Huszár, J. Totz, and Z. Wang, "Real-time
image super-resolution using very deep residual channel attention networks," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 286–301.

T. Shi, J. Caballero, F. Huszár, J. Totz, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in Proc. IEEE Conf. Comput. Pattern Recognition (CVPR), Jun. 2016, pp. 1874–1883.

Y. Guo et al., "Closed-loop matters: Dual regression networks for single image super-resolution," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition, Jun. 2020, pp. 5407–5416.

Y. Chen et al., "Drop an octave: Reducing spatial redundancy in convolutional neural networks," in Proc. IEEE Conf. Comput. Pattern Recognition, Jan. 2021, pp. 10551–10560.

Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution by pursuing reflection image," in Proc. IEEE Conf. Comput. Pattern Recognition (CVPR), Jun. 2021, pp. 10581–10590.

T.-H. Lin, P. Wang, and Y. Wang, "Intrinsic image decomposition with incremental temporal training," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), Jun. 2020, pp. 1629–1639.

J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition (CVPR), Jun. 2021, pp. 2034–2042.

K. Zhang et al., "Deblurring by realistic blurring," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition, Jun. 2020, pp. 2737–2746.

D. Park, U. K. Dong, J. Kim, and S. Y. Chun, “Multi-temporal recurrent neural networks for progressive non-uniform image deblurring with incremental temporal training," in Proc. Eur. Conf. Comput. Vis., 2020, pp. 327–343.

J. Cui, W. Li, W. Guo, and W. Gong, "Progressive downsampling and adaptive guidance networks for dynamic scene deblurring." Pattern Recognition, vol. 132, Dec. 2022, Art. no. 109888.

J. Whang, M. Delbracio, H. Talebi, C. Sahara, A. G. Dimakis, and P. Milanfar, "Deblurring via stochastic refinement," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition (CVPR), Jun. 2022, pp. 16293–16303.

J. Zhang and W. Zhai, “Blind attention geometric restraint neural network for single image dynamic/defocus deblurring.” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 2, 2022.

R. Lu, Y. Yuan, and Q. Wang, “Efficient deblurring via high-frequency and low-frequency information fusion,” in Proc. IEEE Int. Conf. Image Process. (ICIP), Oct. 2022, pp. 1271–1275.

Z. Jiang, Y. Zhang, D. Zou, J. Ren, J. Lv, and Y. Liu, “Learning event-based motion deblurring," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition (CVPR), Jun. 2020, pp. 3320–3329.

M. Siin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR), Jun. 2020, pp. 3606–3615.

Y. Mao, Z. Wan, Y. Dai, and X. Yu, “Deep idempotent network for efficient single image blind deblurring," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 1, pp. 172–185, Jan. 2023.

Q. Zhao, H. Wang, Z. Yue, and D. Meng, “A deep variational Bayesian framework for blind image deblurring," Knowl.-Based Syst., vol. 249, Aug. 2022, Art. no. 109008.

K. Purohit, M. Siin, A. N. Rajagopalan, and V. Naresh Boddeti, “Spatially-adaptive image restoration using distortion-guided networks," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 2300–2319.

D. Huo, A. Masoumzadeh, and Y.-H. Yang, “Blind non-uniform motion deblurring using atrous spatial pyramid deformable convolution and deblurring-reblurring consistency," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition Workshops (CVPRW), Jun. 2022, pp. 437–446.

C. Guo, X. Chen, Y. Chen, and C. Yu, "Multi-stage attentive network for motion deblurring via binary cross-entropy loss," Entropy, vol. 24, no. 10, pp. 1414, Oct. 2022.

S. W. Zamir et al., “Multi-stage progressive image restoration," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition, Jun. 2021, pp. 14821–14831.

H. Lee, H. Choi, K. Sohn, and D. Min, “KNN local attention for image restoration," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition (CVPR), Jun. 2022, pp. 2139–2149.

L. Ruan, B. Chen, J. Li, and M.-L. Lam, “AIFNet: All-in-focus image restoration network using a light field-based dataset,” IEEE Trans. Image Process., vol. 30, pp. 675–688, 2021.

A. Abualaim, M. Affi, and M. S. Brown, “Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning," in Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV), Jan. 2022, pp. 1231–1239.

L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognition (CVPR), Jun. 2022, pp. 16304–16313.
Qiang Li received the B.S. degree from the Department of Computer Science and Technology, Changchun University, China, in 2020. He is currently pursuing the Ph.D. degree with the College of Information Science and Technology, Northeast Normal University, China. His research interests include image restoration.

Miao Qi received the B.S. degree from Anshan Normal University, China, in 2004, the M.S. degree from the College of Computer Science and Information Technology, Northeast Normal University, China, in 2007, and the Ph.D. degree from the College of Chemistry, Northeast Normal University, in 2010. She is currently an Associate Professor with the College of Information Science and Technology, Northeast Normal University. Her research interests include deep learning and pattern recognition.

Di Liu received the B.S. and M.S. degrees from the College of Computer Science and Information Technology, Northeast Normal University, China, in 2006 and 2009, respectively. She is currently pursuing the Ph.D. degree with the College of Information Science and Technology, Northeast Normal University. Her research interests include pattern recognition, machine learning, and image/video understanding.

Jun Kong received the B.S. and M.S. degrees from the Department of Mathematics, Northeast Normal University, China, in 1992 and 1997, respectively, and the Ph.D. degree from the College of Mathematics, Jilin University, in 2001. From 2003 to 2004, he was a Visiting Scholar with Edith Cowan University, Perth, WA, Australia. He is currently a Professor with the College of Information Science and Technology, Northeast Normal University. His research interests include artificial intelligence, digital image processing, pattern recognition, machine learning, biometrics, and information security.

Jianzhong Wang received the Ph.D. degree from the School of Mathematics and Statistics, Northeast Normal University, in 2010. He is currently an Associate Professor with the College of Information Science and Technology, Northeast Normal University. His research interests include pattern recognition and image processing.