Deep Reference-based Dynamic Scene Deblurring

Cunzhe Liu[1], Zhen Hua[1*], and Jinjiang Li[2]
[1] School of Information and Electronic Engineering, Shandong Technology and Business University
Yantai, Shandong 264005, China
[e-mail: goodlucklcz@gmail.com]
[2] School of Computer Science and Technology, Shandong Technology and Business University
Yantai, Shandong 264005, China
[e-mail: huazhen66@foxmail.com]
[*]Corresponding author: Zhen Hua

Received July 10, 2022; revised September 24, 2023; accepted March 12, 2024;
published March 31, 2024

Abstract

Dynamic scene deblurring is a complex computer vision problem owing to its difficulty to model mathematically. In this paper, we present a novel approach for image deblurring with the help of the sharp reference image, which utilizes the reference image for high-quality and high-frequency detail results. To better utilize the clear reference image, we develop an encoder-decoder network and two novel modules are designed to guide the network for better image restoration. The proposed Reference Extraction and Aggregation Module can effectively establish the correspondence between blurry image and reference image and explore the most relevant features for better blur removal and the proposed Spatial Feature Fusion Module enables the encoder to perceive blur information at different spatial scales. In the final, the multi-scale feature maps from the encoder and cascaded Reference Extraction and Aggregation Modules are integrated into the decoder for a global fusion and representation. Extensive quantitative and qualitative experimental results from the different benchmarks show the effectiveness of our proposed method.

Keywords: Image deblurring, reference image, convolutional neural network, encoder-decoder, deep learning.

http://doi.org/10.3837/tiis.2024.03.007 **ISSN : 1976-7277**

1. Introduction

Blind image deblurring is a classic low-level computer vision task, which aims to recover clear images from blurred input images. As more and more photos are taken by hand-held cameras and smart phones, some factors such as camera shake or object movement are unavoidable, and the resulting blurred images are visually unpleasant. Therefore, image deblurring is beneficial to improve the quality of media content so that users can get a better experience. For example, text recognition and photo algorithm for mobile cameras. Besides, this basic technology can also benefit some high-level computer vision tasks, like object recognition [1] and autonomous driving system [2]. Therefore, image deblurring has important application value in both academia and industry.

As in other fields of computer vision, deep learning-based methods have significantly improved the progress of single blind image deblurring. Early deep learning-based single image deblurring works [3,4,5,6,7] mainly focus on kernel estimation and simultaneously recover sharp images with deconvolution. However, the blur of real scenes is complex and the method of blur kernel estimation cannot completely remove the blur. On the other hand, recent deep learning-based methods [8,9,10,11,12,13,14,15,16,17] use deep convolutional neural networks (CNNs) to solve this problem without estimating the blur kernel. Nevertheless, the ill-posed nature of the single-image deblurring problem makes it still difficult for these methods to recover results with structural details, as shown in Fig. 1.

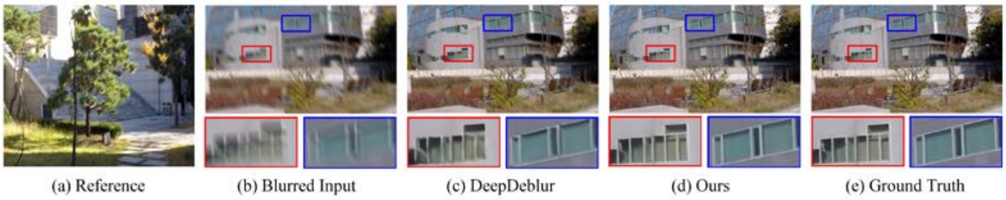

(a) Reference (b) Blurred Input (c) DeepDeblur (d) Ours (e) Ground Truth

Fig. 1. One example. (c) represent the result of DeepDeblur [8]. Our method produces sharper result (d) with the help of the reference image (a). (e) is the ground truth image corresponding to the blurred input (b).

This paper turns from single image deblurring to exploring a new method of image deblurring using clear reference image. We exploit the rich texture of the reference image to compensate for the structural details lost in the blurred image to alleviate this ill-posed problem and produce a clean image with the help of the reference image. Compared with a single image, Clear reference images tend to contain more high-frequency information that can facilitate image deblurring. Clear reference images can be captured using cameras from different angles, or obtained through web image searches, etc. The reference based approach is widely used in low-dimensional vision tasks, particularly in the field of Reference-based Super-Resolution(RefSR) [18,19,20,21,22,23]. As a promising work, Wang et al. [22] explored dual-camera super-resolution as a RefSR problem, leveraging the camera's telephoto photos to improve the resolution of wide-angle photos. However, due to the different processing mechanisms for generating low-resolution images and blurred images, it is not effective to directly apply RefSR approach on image deblurring, and only a slight deblurring effect can be achieved, which is far from the baseline level. Moreover, these RefSR methods usually require input images and reference images to have similar content, which greatly limits the use of clear reference pictures. Until now, there has been little exploration of image deblurring with reference images [24] [25], a work that is closely related to ours is HaCohen et al. [25], who used reference images for kernel estimation and strong local priors for the non-blind deconvolution step. Although their method can generate deblurred images with less artifacts, when the content of the image is in a region with large differences, the similar content of the blurred image and the reference image cannot be matched, and when the estimated blur kernel is not accurate enough, the performance is limited when dealing with complex blurred images. Therefore, their method is not practical in complex scenes with object motion and camera shake. In essence, the key to the success of our work is to establish the correspondence between the blurred image and the reference image and effectively transfer the most relevant texture details.

To address the above limitations, we propose a reference-based deblurring method, which aims to explore the most relevant features from the reference image to help remove the blur. Our network is an encoder-decoder based network architecture. More specifically, our network mainly includes encoders, Spatial Feature Fusion Module, Reference Extraction and Aggregation Module, and fusion decoder. The encoders are used to extract the multi-scale features of the blurred input and reference images. The spatial feature fusion module is introduced in the process of blurred input encoding, which combines downsized feature and downsampled blurry input feature to facilitate the process of deep feature extraction. As for the reference extraction and aggregation module, it is developed to search and transfer the most relevant texture detail information from the reference features to input features. The fusion decoder is developed to aggregates feature maps from both domains to restore the latent deblurred image. Experimental results demonstrate that our method surpasses other single image deblurring algorithms on public datasets and it is also robust to different reference images.

The main contributions of our paper are:

- We propose an end-to-end trainable reference-based deblurring method that utilizes the most relevant features from the reference image to assist the image deblurring.
- As far as we know, there are few studies exploring reference-based image deblurring, and our method breaks the conventional idea of using only one image for image deblurring. Even with clear reference images of other scenes, our method can recover clear images with better structural details.
- We introduce the Spatial Feature Fusion Module in the encoder and integrate the multi-scale feature maps from the encoder and cascaded Reference Extraction and Aggregation Module in the decoder. In this way, not only the spatial image detail is preserved in the encoder, but also the high-frequency information from the reference image is fused in the decoder.
- We conduct quantitative and qualitative evaluations on three public datasets, and experimental results demonstrate that our method effectively promotes image deblurring performance.

2. Related Work

During this section, we briefly review some of the previous research relevant to our research, including deep image deblurring and reference-based methods.

2.1 Deep Image Deblurring

In recent years, deep learning methods have made significant progress in image deblurring tasks. To achieve excellent performance, many researchers use deep learning methods to build end-to-end models. Sun et al. [3] proposed a CNN-based model to remove inconsistent motion blur by estimating blur kernels. Since kernel estimation is usually sensitive to noise, inaccurate kernel estimation can lead to poor image restoration. The majority of deep learning-based methods move to predict sharp images from blurry images using direct image-to-image regression. These promising approaches broadly employ three strategies: multi-scale(MS), multi-patch(MP), and multi-temporal(MT) methods. Nah et al. [8] proposed a MS convolutional neural network in a coarse-to-fine manner without estimating the blur kernel and using the coarse-scale image to gradually restore the fine-scale clear image. Since this network does not share parameters among different scales, this leads to abundant parameters and high calculation costs. In order to solve this problem, Tao et al. [9] proposed a Scale-recurrent Network based on the strategy of MS, which shares weight parameters between different scales, has less model parameters, is less computationally expensive and Excellent performance is achieved. Zhang et al. [11] utilize an MP method and develop a hierarchical deblurring network. Park et al. [14] proposed a new MT method. By constructing datasets with different Temporal levels (TL), iteratively trained progressively from blurry images with higher TL to adjacent images with low TL until TL is 1. Li et al. [17] proposed SimpleNet with an encoding-decoding architecture, which enhanced the image details using tensor decomposition theory, and repeatedly placing atrous convolutions to enhance the receptive field of the entire network, achieving comparable performance. Moreover, Kupyn et al. [12] proposed DeblurGAN based on generative adversarial network, which significantly improved the subjective visual quality. However, the specific texture structure in the restored image is still difficult to restore, and there is still a lot of room for improvement. On the basis of the DeblurGAN algorithm, Kupyn et al. [13] proposed the DeblurGAN-v2 algorithm with a more advanced deblurring effect, which uses the spatial pyramid network as the core module to build a generator. At the same time, networks of different magnitudes can also be selected as the backbone network of the spatial pyramid, and a good restoration effect has been achieved. These end-to-end methods only use blurred images as input, so it is difficult to recover deblurred images with structural details. Instead, we introduce additional clear reference images to help the network recover deblurred images with more details and structure.

2.2 Reference-based Methods

Compared with the deblurring algorithm using only a single blurred image as input, the reference-based method introduces additional clear images, which can provide more image information. Reference-based methods are helpful for processing low-level vision tasks, especially in the super-resolution field. Reference-based super-resolution removes the ill-posed essence of single image super-resolution by exploiting the high-frequency details of the reference image. For establishing the correspondence between the reference image and the input image, optical flow or patch-based matching are employed. Specifically, Zheng et al. [19] proposed an encoder-decoder network that uses optical flow to align input image and reference image and then fuses the multi-scale features of the warp into the decoder accordingly. However, optical flow estimation still suffers a large performance drop in unaligned regions. Yang et al. [21] adopted the learnable patch-match method to introduce the transformer into the reference-based image super-resolution task, and transfer the most relevant texture features by computing the hard attention and soft attention between the input image and the reference image. Recently, Lu et al. [23] used a course-to-fine corresponding matching mechanism to extract the features with the highest similarity between feature patches under the strategy of patch matching, which greatly reduced the computational cost. Wang et al. [22] applied the reference-based method to the images captured by different focal length lenses of smartphones, taking the image obtained by the telephoto lens as the reference image, extracting reference features of different scales, and applying aligned attention to warp it for matching the input and performing super-resolve on the low-resolution image obtained by the wide-angle lens to obtain high-quality and high-fidelity results.

These methods mentioned above demonstrate the importance of reference images in low-level visual image enhancement tasks. However, these methods require the reference images and input images to have similar content or good alignment, which limits the availability and effectiveness of clear reference images. In contrast, in our experiments, the reference image does not need to have similar content and better alignment with the blurred image. Instead, we only find and transfer the most relevant features from the reference image to help the restoration of blurred images, resulting in clean images with structural details.

3. Method

Since U-Net [26] has shown excellent performance in the field of image deblurring [27], we adopt it as the basic architecture of our network. As shown in **Fig. 2**, our proposed network consisting three parts: encoders for feature extraction, Reference Extraction and Aggregation Module (REAM), Spatial Feature Fusion Module (SFFM), and decoder for feature reconstruction. I_{Blur} and I_{Ref} represent the blurred input and the reference image, respectively. To take full advantage of the structural details of the reference image, we first use a shallow encoder to extract features at different scales of the reference image. To be specific, the shallow encoder consists of three groups convolution layers and the second and third groups halve the size of the feature maps with stride 2. After passing through the shallow encoder, we get three reference features at different scales and can be expressed as F_{Ref}^{n}, where n = 1, 2, 3. Afterwards, I_{Blur}, I_{Ref} and F_{Ref}^{n} are fed into REAM to perform feature matching extraction and aggregation as shown in **Fig. 3**. Before performing feature extraction, we first perform a matching operation between I_{Blur} and I_{Ref}. The feature extraction operation is performed three times, each for one reference feature F_{Ref}^{n} of scale s. At the same time, we perform deep feature extraction on the blurred input image, we use a convolutional layer and three ResBlock groups (RGs) to extract multi-scale feature maps and place SFFM in front of the last two RGs to preserve spatial image information. Each RG consists of multiple ResBlocks. ResBlock consists of two 3×3 convolutional layers and a Rectified Linear Unit (ReLU). To obtain the final deblurred output, we concatenate the same scale features from the encoder (if any) and REAM in the decoder. Lastly, our pipeline ends with a skip connection to the blurred input image.

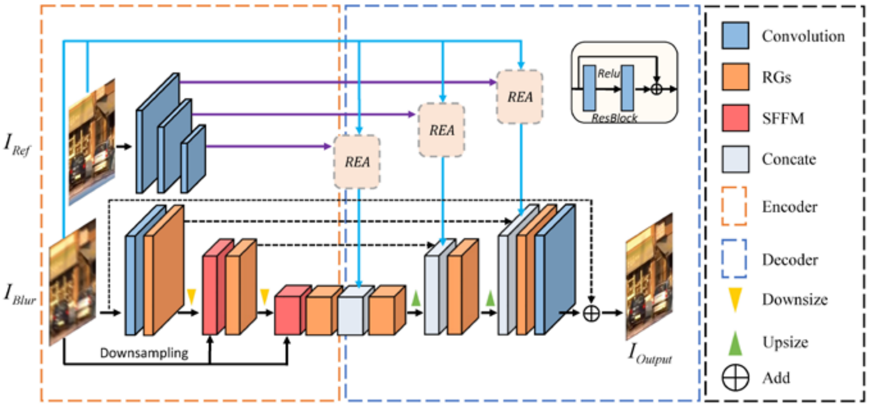


Fig. 2. Network structure of our approach. The “Downsize” is a convolution with stride 2; the “Upsize” is a transposed convolution with stride 2.

3.1 Reference Extraction and Aggregation Module(REAM)

To transfer the reliable high quality features from the clear reference image, we use three Reference Extraction and Aggregation Modules(REAM) between the encoder and decoder. For the sake of clarity, we will use the last REAM as an example for a detailed introduction.

As shown in **Fig. 3**, our REAM consists of three parts: matching, extraction and aggregation process. The matching part calculates the cosine similarity matrix between I_{Blur} and I_{Ref} and computes the index map P and confidence map C for use in the next two parts. The extraction part uses the index map to extract the most relevant features from the reference features. The final aggregation process fuses the reference features with the input features guided by the matching confidence map. Next, we will introduce each part in detail.

Matching: We use a shared encoder ϕ [28] to map I_{Blur} and $I_{Ref}\downarrow\uparrow$ to the feature space, where $\downarrow$ and $\uparrow$ denotes the down-sampling operator and up-sampling operator respectively. Note that the repeated downsampling and upsampling of the reference image is to obtain a blurred reference image, reducing the domain gap between the blurred input image and the reference image to facilitate matching between them. Then we compute the cosine similarity matrix between the blurred features and the reference features, specifically, we use a 3×3 sliding window to unfold the blurred and reference features into N non-overlapping patches respectively: $\{P^{0}_{Blur,...,}P^{N-1}_{Blur}\}$, $\{P^{0}_{Ref,...,}P^{N-1}_{Ref}\}$. We further do a dense patch matching operation, for each patch in P^{N}_{Blur} and each patch in P^{N}_{Ref}, we compute the cosine similarity matrix R between them:

$$R_{i,j}=\left\langle\frac{q_i}{\|q_i\|},\frac{p_i}{\|p_i\|}\right\rangle \tag{1}$$

where $R_{i,j}$ is the similarity between the i-th patch of the blurred feature and the j-th patch of the reference feature. For each blurred feature patch, our goal is to find its most relevant reference feature patch for later feature aggregation. So we calculate the matching index map and the confidence map through the argmax operation and the max operation respectively. This process can be calculated as:

$$P_i=\arg\max_j r_{i,j} \tag{2}$$

$$C_i=\max_j r_{i,j} \tag{3}$$

where P_i represents the patch index of the reference feature, which is most relevant to the feature of the i-th patch of the blurred feature, and C_i provides the matching confidence for the subsequent feature transfer.

Extracting: After obtaining the index map, we can extract the most relevant features according to the index, but before feature extraction, we need to unfold the reference features into patches, which can be expressed as: $\left\{B^{0}_{Ref},...,B^{N-1}_{Ref}\right\}$. Then the clear reference features B^{N}_{Ref} and matching index P_i are fed into the extraction module to extract the most relevant reference features, and this process can be expressed as:

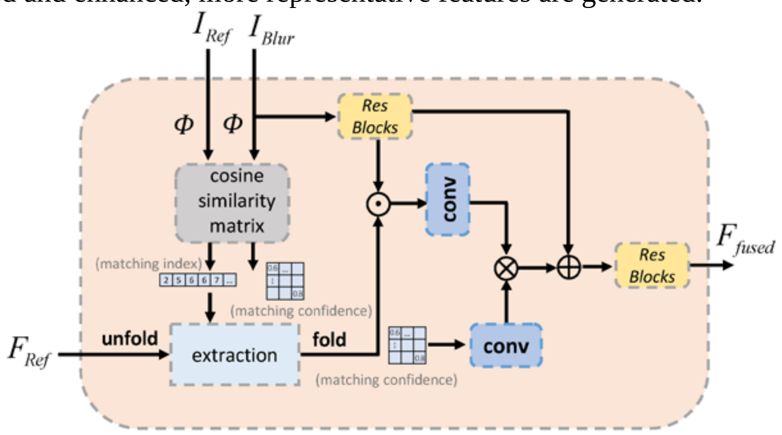
$$B^{N}_{M}=E(B^{N}_{Ref}\mid P_i) \tag{4}$$

where $E(\cdot)$ represents the extraction module, B^{N}_{M} is the most relevant features obtained from the reference features according to the index map P_i. we transfer feature patches according to the most relevant position. After obtaining a new feature map B_M^N , we need to fold $\{B_M^0, ..., B_M^{N-1}\}$ to restore to the original feature dimension, denoted as B_F . Folding is the inverse operation of unfolding.

Aggregation Process: In this aggregation process, we fuse the most relevant reference features obtained with the input features. We first used residual blocks to map I_{Blur} to the feature space, denoted as: F_{Blur} , then concatenate with the most relevant reference feature B_F , Simple addition or concatenation will bring additional noise or irrelevant information. Inspired by [23], we use matching confidence for adaptive spatial fusion, which only weights the most relevant features. The aggregation process can be expressed as:

$$F_{fused} = \{conv([F_{blur}, B_F]) \otimes conv(C)\} + F_{blur} \tag{5}$$

where [,] an $\otimes$ indicate concatenation operation and element-wise multiplication, respectively. As a result, the significant information in the input features and reference features are aggregated and enhanced, more representative features are generated.



(a) Reference Extraction and Aggregation Module

Fig. 3. Illustration of the Reference Extraction and Aggregation Module.

3.2 Spatial Feature Fusion Module(SFFM)

The design of the network architecture for the deblurring task requires a large receptive field to deal with more serious motion blur [9]. To this end, we have carried out relatively deep feature extraction by stacking Resblock groups in the encoder. However, it is difficult to preserve reliable spatial detail information due to the gradually shrinking feature map in the process of deep feature extraction, which is not conducive to image reconstruction and restoration. Inspired by [29,30], we design Spatial Feature Fusion Module(SFFM) to compensate for the loss of spatial information during feature extraction, instead of simply concatenating the downsampled original blurry image to the encoder.

As shown in **Fig. 4**, we first extract features from the downsampling blurred input using four 3×3 convolutional layers , this feature is then concatenated with the original input blurred image, and further refined with a 1×1 convolutional layer to obtain the parameter γ of the same size as the downsized feature. Then γ is fed into another convolutional layer to get the parameter β , and finally, the downsized feature F multiplied λ and added to β in an element-wise manner as：

$$SFF(F \mid \gamma, \beta) = \gamma \otimes F + \beta \tag{6}$$

where $\otimes$ denotes the element-wise multiplication. We use the learnable convolution layer to predict parameters λ and β, and perform element transformation on the input feature F. By using SFFM, the encoder effectively combines the spatial information from the different scales in the process of deep feature extraction and enables the encoder to perceive the blur information of different scales, which promotes the process of deep feature extraction.

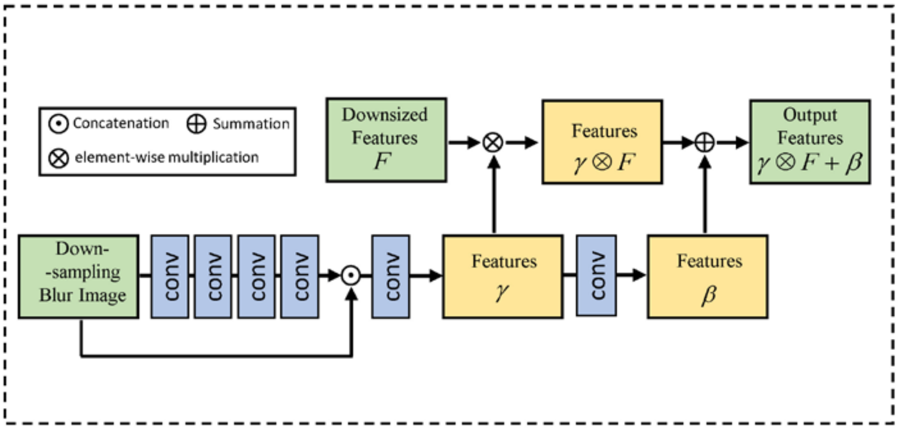


Fig. 4. Illustration of the Spatial Feature Fusion Module.

3.3 Loss Functions

The loss function is very important because it optimizes our model parameters until we find the best model parameters. To recover deblurred images with good visual quality and high-frequency structural information, our network is constrained by the following loss function:

$$\mathcal{L}_{total} = \lambda_{con}\mathcal{L}_{con} + \lambda_{freq}\mathcal{L}_{freq} + \lambda_{adv}\mathcal{L}_{adv} \tag{7}$$

where $\mathcal{L}_{con}$, $\mathcal{L}_{freq}$ and $\mathcal{L}_{adv}$ represent the content loss, frequency reconstruction loss and adversarial loss, respectively. λ_{con}, λ_{freq} and λ_{adv} represent the weights of content loss, frequency loss and adversarial loss.

Content loss usually uses L1 loss or MSE loss, here we use L1 loss because it performs better in our experiments. L1 loss can be defined as:

$$\mathcal{L}_{con} = \left\| I_{DE} - I_{GT} \right\|_1 \tag{8}$$

where I_{DE} represents the model output and I_{GT} represents the corresponding ground truth image.

Frequency reconstruction loss [27] recovers more high-frequency details by minimizing the difference between deblurred images and ground truth image in the frequency domain, facilitating the generation of results with structural details. Different from [27], we adopt a single-scale frequency reconstruction loss function:

$$\mathcal{L}_{freq} = \left\| \mathcal{F}\left(I_{DE}\right) - \mathcal{F}\left(I_{GT}\right) \right\|_2 \tag{9}$$

where F is Fourier transform.

Adversarial loss [31] can produce better visual quality results, therefore, we chose Relativistic GANs [32]:

$$\mathcal{L}_D = -\mathbb{E}_{I_{GT}}\left[\log(D(I_{GT},I_{DE})\right] - \mathbb{E}_{I_{DE}}\left[\log(1-D(I_{DE},I_{GT})\right] \tag{10}$$

$$\mathcal{L}_G = -\mathbb{E}_{I_{GT}}\left[\log(1-D(I_{GT},I_{DE})\right] - \mathbb{E}_{I_{DE}}\left[\log(D(I_{DE},I_{GT})\right] \tag{11}$$

4. Experiments

In this section, we evaluate our model. We first present the relevant experimental details. Our experimental results are then compared with state-of-the-art methods quantitatively and qualitatively. Lastly, we implement ablation experiments to verify the validity of our proposed method.

4.1 Dataset and Implementation Details

For training and testing our method, we use the widely adopted GoPro [8] synthetic dataset, obtained at 240 frames rate using a GoPro4 Hero Black high-speed camera. The blurred images are synthesized by averaging neighboring frames, and the intermediate frames are used as ground truth image. It contains 2103 pairs of blurred and sharp images for training and 1111 pairs for testing. For the selection of clear reference images, considering that the GoPro dataset can be viewed as a continuous video frame [14,17], we choose nearby frames of sharp images as reference images, rather than the ground truth image itself. This is because in real life, the same shooting scene is difficult to reproduce. In addition, we also use only the HIDE [33] test dataset to evaluate the generalization ability of our model, which contains 2025 pair of test images, where the reference images are selected from the adjacent images. For the performance of handling real-world blur, we conduct experiments on RealBlur-J [34], which contains 3758 pairs of training images and 980 pairs of test images, whose reference images are selected from the adjacent clear folder. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Measure (SSIM) [35] are used as evaluation metrics.

The batch size selected for our network training is 12. We choose Adam as the optimizer and set the initial learning rate to 10^{-4} and apply the cosine annealing strategy to decrease steadily. We randomly crop the input size to 120×120 and the activation function uses the ReLU. The hyperparameters of the loss function λ_{con}, λ_{freq} and λ_{adv} are 1, 0.1 and 5×10^{-3}, respectively. We use a server with eight NVIDIA TITAN RTX GPUs for experiments and implement the model under the framework of PyTorch 1.6 [36]. Our model contains six RGs, each of which consists of n modified ResBlocks [9]. Properly increasing the size of n can enhance the deblurring performance of the network. We have applied several different numbers of ResBlocks. The larger the value of n, the larger the corresponding parameter amount. As shown in **Table 1**, when n is equal to 16, the PSNR score is 32.50 and the parameter amount is 14.09. Our model achieves a good balance between model parameters and performance. Therefore, we choose to place 16 ResBlocks in each RG.

Table 1. Evaluation of different numbers of Resblocks in Resblock group. Due to training time and GPU memory limitations, we did not try a larger number n.

N	PSNR	SSIM	Params
8	31.71	0.951	7.59
12	32.25	0.955	10.99
16	32.50	0.958	14.09
20	32.19	0.955	17.19

4.2 Experimental Results

Quantitative evaluations. We compare the experimental metrics on the GoPro test dataset with some of the most advanced methods [8,9,10,11,13,14,15,16,17]. **Table 2** shows the PSNR and SSIM scores, our method is higher than any other single image deblurring method.

Our method tightly grasps the essence that the reference image has high-frequency information and clear texture, and effectively transfers the most relevant texture detail information to the deblurred image. **Table 3** and **Table 4** demonstrate the evaluation results on HIDE and RealBlur-J test datasets, respectively. Once again, our method achieves the best results compared to other algorithms. This shows that our algorithm has more reliable deblurring performance than state-of-the-art methods, both on synthetic datasets and in the real scenes.

Table 2. Evaluation results on the benchmark GoPro testing set. * represents the author did not release their codes. The two best scores are marked red and orange in each column, respectively.

Method	**PSNR(dB)**	**SSIM**	**Params(M)**
DeepDeblur [8]	27.83	0.915	11.7
SRN [9]	30.25	0.935	6.8
DeblurGAN-v2 [13]	29.55	0.934	-
MTRNN [14]	31.13	0.945	2.6
DSD [10]	30.96	0.942	2.84
Stack4-DMPHN [11]	31.39	0.948	21.7
RADN* [15]	31.76	0.953	-
SAPHN* [16]	32.02	0.953	-
SimpleNet* [17]	31.52	0.950	-
Ours	32.50	0.958	14.09

Table 3. Evaluation results on the benchmark HIDE dataset. Note that all the models are trained using only GoPro training dataset. The two best scores are marked red and orange in each column, respectively.

Method	**PSNR(dB)**	**SSIM**	**Params(M)**
DeepDeblur [8]	25.73	0.874	11.7
DeblurGAN-v2 [13]	27.40	0.882	-
SRN [9]	28.36	0.904	6.8
DSD [10]	29.01	0.913	2.84
Stack4-DMPHN [11]	29.10	0.918	21.7
MTRNN [14]	29.15	0.918	2.6
SAPHN* [16]	29.98	0.930	-
Ours	30.57	0.935	14.09

Table 4. Evaluation results on the RealBlur-J testing set. The two best scores are marked red and orange in each column, respectively.

Method	**PSNR(dB)**	**SSIM**	**Params(M)**
DeblurGAN-v2 [13]	29.69	0.870	-
SRN [9]	31.38	0.909	6.8
Ours	31.55	0.912	14.09

Qualitative evaluations. We next carry out a visual quality assessment. Considering the pioneering nature of DeepDeblur [8] and the excellent performance of SRN [9], DeblurGAN-v2 [13], DMPHN [11] and MTRNN [14], we selected these five deblurring methods for qualitative comparison. **Fig. 5** and **Fig. 6** demonstrate the visual comparison results of our method with these methods on the GoPro and the HIDE test dataset. The first example of the **Fig. 5** is a shooting scene with low-speed motion, while the second and third examples are shooting scenes with high-speed motion. It can be observed that our model achieves the best deblurring effect while other methods are more difficult to remove severe motion blur. For example, the license plate number of the first example, with the help of the reference image, our method recovers the deblurred image with more structural details. In two other examples of high-speed motion blur, our method can also reconstruct sharper vehicle outlines and recognizable text. In **Fig. 6**, our method can restore more realistic facial expressions and text textures on the HIDE test dataset while other methods lose more image details. For example, in the second example, the facial features produced by other methods are still very blurred. In contrast, our method produces results with finer details on the boy's eyebrows, nose, and eyes.

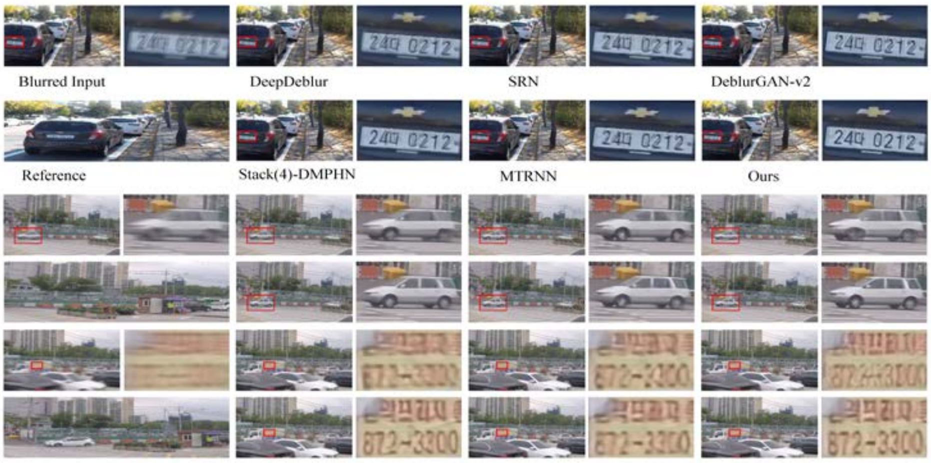


Fig. 5. Qualitative comparisons on GoPro test dataset. Our method restores clearer detail texture on texts and object.

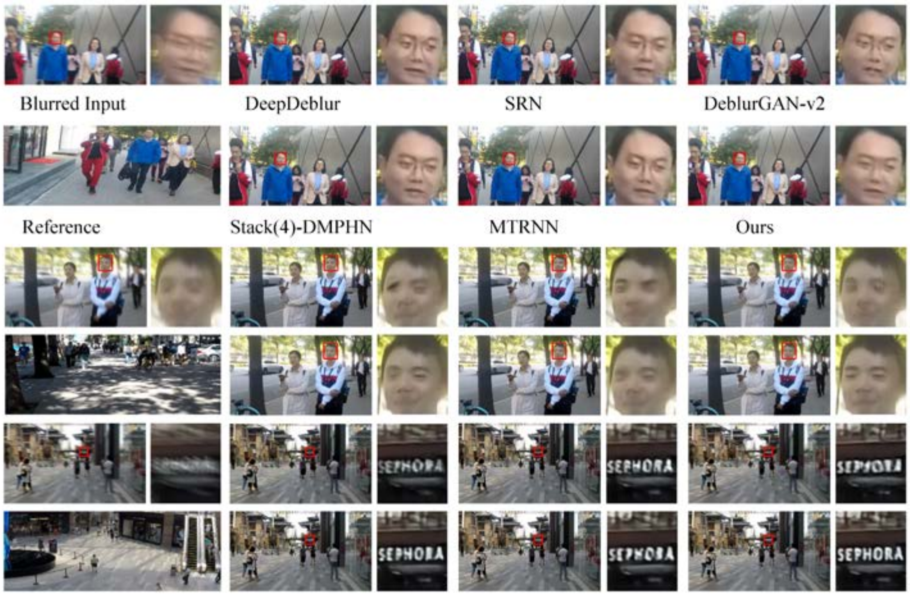


Fig. 6. Qualitative comparisons on HIDE test dataset. Our method restores clearer detail texture on face and texts.

In **Fig. 7**, we demonstrate the qualitative comparison results of our method with SRN and DeblurGAN-v2 on RealBlur-J. This dataset contains real-world blur in low-light situations. As can be seen, other methods still have blurred structures, while our method recovers sharp details. All in all, the quantitative assessment results show that our algorithm has excellent performance both on synthetic datasets and real-world scenarios.

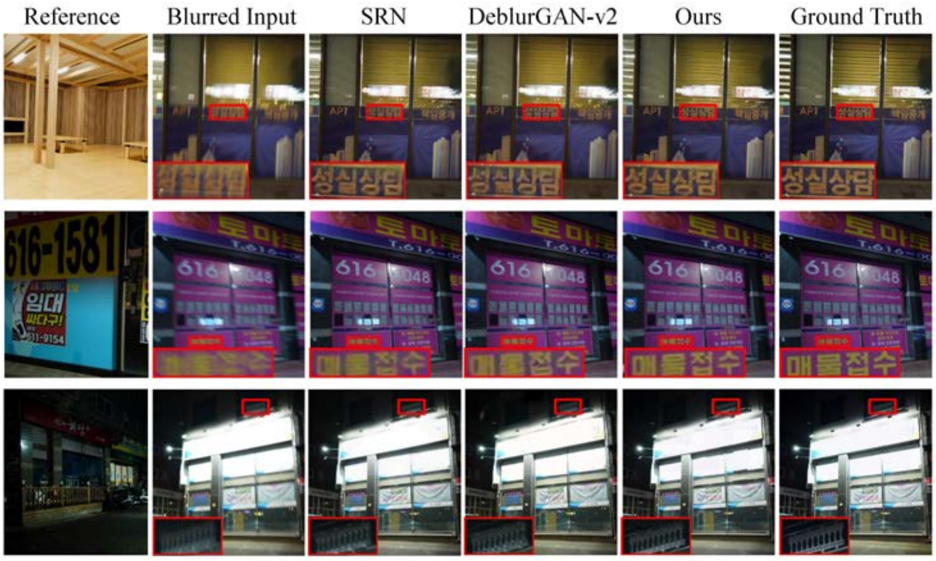


Fig. 7. Qualitative comparisons on RealBlur-J test dataset. Our method can still restore the realistic structural details in low light.

4.3 Ablation Study

In the previous comparison, we demonstrate the excellence of our approach, below we will conduct ablation experiments on components to further investigate our model.

4.3.1 Effective of Spatial Feature Fusion Module

The SFFM plays a role in compensating spatial information in the initial stage of feature extraction. To confirm the validity of SFFM, we replace it with vanilla Resblocks. As shown in **Table 5**, for PSNR and SSIM, SFFM obtains gains of 0.23 dB and 0.002.

Table 5. Impact of the SFFM on the performance of our model on GoPro dataset. The best score is marked red.

	Without SFFM	**Ours(with SFFM)**
PSNR(dB)	32.27	32.50
SSIM	0.955	0.958

Fig. 8 demonstrates the results of the qualitative comparison, and the deblurring results using SSFM have higher quality results.

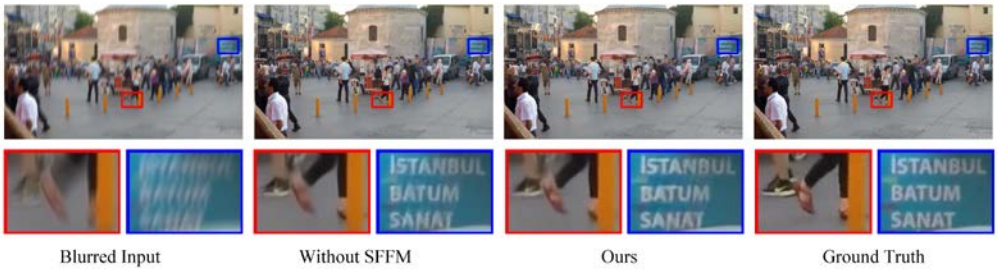


Fig. 8. Ablation study on the spatial feature fusion module.

4.3.2 Effective of the Reference Image

Our method relies on a clear reference image to recover deblurred images with texture details and structural information. An unavoidable question is whether the reference image has any substantial help in image restoration. To confirm the validity of the reference image, we perform an experiment that removes all modules related to the reference image and keep the encoder-decoder structure and other components unchanged and then retrain on the GoPro dataset. **Table 6** and **Fig. 9** show our quantitative and qualitative analysis results, respectively. From **Table 6** we can observe that compared with the model without the help of the reference image, the PSNR of our method is 0.38 dB higher, which verifies the effectiveness of the reference image for image deblurring of our proposed method.

Table 6. Impact of the reference image on the performance. The best score is marked red.

	Without Reference	**Ours(with reference)**
PSNR(dB)	32.12	32.50
SSIM	0.955	0.958

The results of the visual comparison are shown in **Fig. 9**, the results without using the reference image are still blurred, on the contrary, our deblurred results have clear structural details. Both quantitative and qualitative analyses above indicate the usefulness of reference images in our approach.

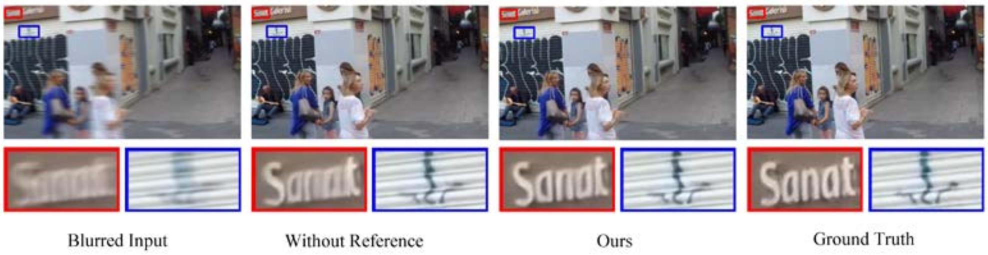


Fig. 9. Ablation study on the reference image.

4.3.3 Effective of Different Reference Images

In the previous subsection, we showed the importance of clear reference images to our method. Another question is whether our proposed image deblurring method is only effective with the help of a unique reference image. To this end, we conduct experiments on the GoPro dataset using randomly different reference images for blurred inputs. As shown in **Fig. 10**, **Fig. 10**(f), (g) and (h) represent the results of deblurring using (b), (c) and (d) as reference images respectively. Note that we randomly selected the reference images from the GoPro dataset. As shown in **Fig. 10**, the deblurred visual results produced by using different reference images are quite similar.

In addition, we also perform quantitative comparisons on blurred input with randomly selected different reference images on the GoPro dataset. **Table 7** shows the quantitative analysis results using different reference images and the two metrics are only slightly different. The above quantitative and qualitative analysis shows that our method is robust even when using reference images of different scenes.

Table 7. Impact of the different reference images on the performance

	Reference1	Reference2	Reference3
PSNR	32.51	32.49	32.49
SSIM	0.956	0.956	0.956

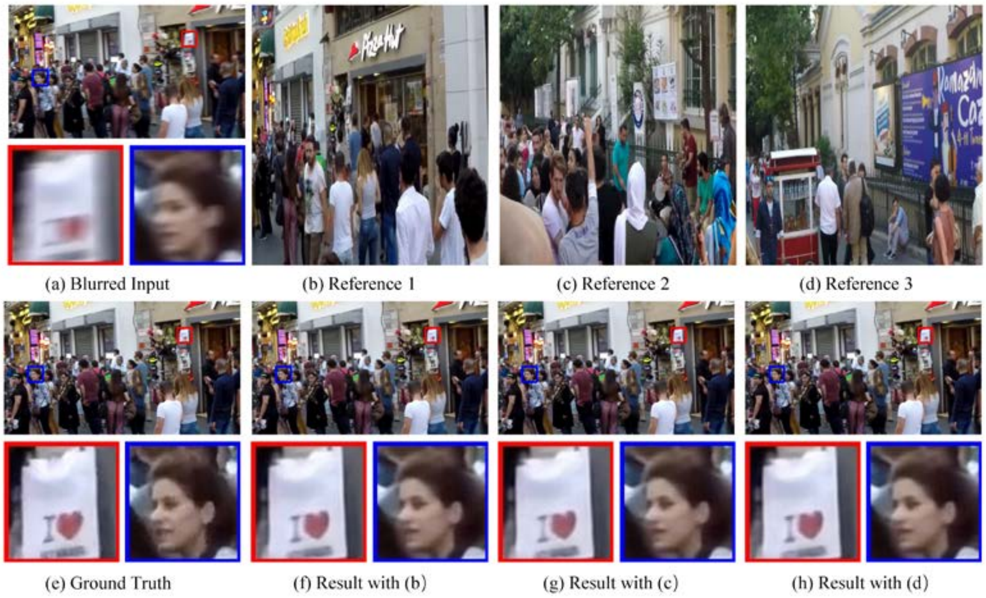


Fig. 10. Ablation study on the different reference images.

5. Conclusion

In this work, we introduce reference-based dynamic scene deblurring. Our network is a fully convolutional trainable encoder-decoder architecture. The proposed Reference Extraction and Aggregation Module effectively finds the most relevant features to facilitate image deblurring and is also robust to different reference images. Further, we also design a novel Spatial Feature Fusion Module to perceive blur at different spatial scales, making up for the lost spatial information. Experiments on prevalent and real-world datasets show the effectiveness of our method and prove that our model is robust to different reference images.

Limitations and Future Work. Since a large amount of memory is occupied during the global matching of blurred images and reference images, we next plan to improve the matching scheme to reduce the memory cost. In future work, it will be a fascinating work to improve the inference speed of the algorithm and reduce the model size for deployment to mobile devices.

References

[1] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin and J. Matas, "DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8183-8192, 2018. Article (CrossRef Link)

[2] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu and H. Li, "Adversarial Spatio-Temporal Learning for Video Deblurring," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 291-301, Jan. 2019. Article (CrossRef Link)

[3] J. Sun, Wenfei Cao, Zongben Xu and J. Ponce, "Learning a convolutional neural network for non-uniform motion blur removal," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 769-777, 2015. Article (CrossRef Link)

[4] C. J. Schuler, M. Hirsch, S. Harmeling and B. Schölkopf, "Learning to Deblur," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 7, pp. 1439-1451, 1 July 2016. Article (CrossRef Link)

[5] Chakrabarti A, "A neural approach to blind motion deblurring," in *Proc. of European conference on computer vision*, pp. 221-235, 2016. Article (CrossRef Link)

[6] X. Xu, J. Pan, Y. Zhang and M. Yang, "Motion Blur Kernel Estimation via Deep Learning," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 194-205, Jan. 2018. Article (CrossRef Link)

[7] Michal Hradiš, Jan Kotera, Pavel Zemčík and Filip Šroubek, "Convolutional Neural Networks for Direct Text Deblurring," in *Proc. of British Machine Vision Conference*, pp. 6.1-6.13, 2015. Article (CrossRef Link)

[8] S. Nah, T. H. Kim and K. M. Lee, "Deep Multi-scale Convolutional Neural Network for Dynamic Scene Deblurring," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 257-265, 2017. Article (CrossRef Link)

[9] X. Tao, H. Gao, X. Shen, J. Wang and J. Jia, "Scale-Recurrent Network for Deep Image Deblurring," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8174-8182, 2018. Article (CrossRef Link)

[10] H. Gao, X. Tao, X. Shen and J. Jia, "Dynamic Scene Deblurring With Parameter Selective Sharing and Nested Skip Connections," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3843-3851, 2019. Article (CrossRef Link)

[11] H. Zhang, Y. Dai, H. Li and P. Koniusz, "Deep Stacked Hierarchical Multi-Patch Network for Image Deblurring," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5971-5979, 2019. Article (CrossRef Link)

[12] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin and J. Matas, "DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8183-8192, 2018. Article (CrossRef Link)

[13] O. Kupyn, T. Martyniuk, J. Wu and Z. Wang, "DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better," in *Proc. of IEEE/CVF International Conference on Computer Vision*, pp. 8877-8886, 2019. Article (CrossRef Link)

[14] D. Park, D. U. Kang, J. Kim and S. Y. Chun, "Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in *Proc. of European Conference on Computer Vision*, pp. 327-343, 2020. Article (CrossRef Link)

[15] K. Purohit and A. N. Rajagopalan, "Region-Adaptive Dense Network for Efficient Motion Deblurring," in *Proc. of AAAI Conference on Artificial Intelligence*, vol. 34, no. 7, pp. 11882-11889, 2020. Article (CrossRef Link)

[16] M. Suin, K. Purohit and A. N. Rajagopalan, "Spatially-Attentive Patch-Hierarchical Network for Adaptive Motion Deblurring," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3603-3612, 2020. Article (CrossRef Link)

[17] J. Li, W. Tan and B. Yan, "Perceptual Variousness Motion Deblurring with Light Global Context Refinement," in *Proc. of IEEE/CVF International Conference on Computer Vision*, pp. 4096-4105, 2021. Article (CrossRef Link)

[18] Haitian Zheng, Mengqi Ji, Lei Han, Ziwei Xu, Haoqian Wang, Yebin Liu and Lu Fang, "Learning Cross-scale Correspondence and Patch-based Synthesis for Reference-based Super-Resolution," in *Proc. of British Machine Vision Conference*, pp. 138.1-138.13, 2017. Article (CrossRef Link)

[19] H. Zheng, M. Ji, H. Wang, Y, Liu and L. Fang, "Crossnet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proc. of the European conference on computer vision*, pp. 87-104, 2018. Article (CrossRef Link)

[20] Z. Zhang, Z. Wang, Z. Lin and H. Qi, "Image Super-Resolution by Neural Texture Transfer," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7974-7983, 2019. Article (CrossRef Link)

[21] F. Yang, H. Yang, J. Fu, H. Lu and B. Guo, "Learning Texture Transformer Network for Image Super-Resolution," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5790-5799, 2020. Article (CrossRef Link)

[22] T. Wang, J. Xie, W. Sun, Q. Yan and Q. Chen, "Dual-Camera Super-Resolution with Aligned Attention Modules," in *Proc. of IEEE/CVF International Conference on Computer Vision*, pp. 1981-1990, 2021. Article (CrossRef Link)

[23] L. Lu, W. Li, X. Tao, J. Lu and J. Jia, "MASA-SR: Matching Acceleration and Spatial Adaptation for Reference-Based Image Super-Resolution," in *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6364-6373, 2021. Article (CrossRef Link)

[24] Y. Hacohen, E. Shechtman, D. Goldman and D. Lischinski, "Non-rigid dense correspondence with applications for image enhancement," *ACM transactions on graphics*, vol. 30, no. 70, pp. 1-10, July. 2011. Article (CrossRef Link)

[25] Y. Hacohen, E. Shechtman and D. Lischinski, "Deblurring by Example Using Dense Correspondence," in *Proc. of IEEE International Conference on Computer Vision*, pp. 2384-2391, 2013. Article (CrossRef Link)

[26] O. Ronneberger, P. Fischer and T Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. of International Conference on Medical image computing and computer-assisted intervention*, pp. 234-241, 2015. Article (CrossRef Link)

[27] 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. of IEEE/CVF International Conference on Computer Vision*, pp. 4621-4630, 2021. Article (CrossRef Link)

[28] K. Simonyan and Z. Andrew, "Very deep convolutional networks for large-scale image recognition," in *Proc. of International Conference on Learning Representations*, 2014. Article (CrossRef Link)

[29] T. Zhang, J. Li amd H. Fan, "Progressive edge-sensing dynamic scene deblurring," *Comp. Visual Media*, vol. 8 no. 3, pp. 495-508, September. 2022. Article (CrossRef Link)

[30] Y. Chi, J. Li and H. Fan, "Pyramid-attention based multi-scale feature fusion network for multispectral pan-sharpening," *Applied Intelligence*, vol. 52, no. 5, pp. 5353-5365, 2022. Article (CrossRef Link)

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets, " *Advances in neural information processing systems*, vol. 27, 2014. Article (CrossRef Link)

[32] A. Jolicoeur-Martineau, "The relativistic discriminator: a key element missing from standard GAN, " in *Proc. of International Conference on Learning Representations*, 2019. Article (CrossRef Link)

[33] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu and L. Shaoet, "Human-Aware Motion Deblurring," in *Proc. of IEEE/CVF International Conference on Computer Vision*, pp. 5571-5580, 2019. Article (CrossRef Link)

[34] J. Rim, H. Lee, J. Won, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *Proc. of European Conference on Computer Vision*, pp. 184-201, 2020. Article (CrossRef Link)

[35] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004. Article (CrossRef Link)

[36] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. of International Conference on Neural Information Processing Systems*, no.721, pp. 8026–8037, December. 2019. Article (CrossRef Link)

Cunzhe Liu received his B.S. degree in Computer Science and Technology from Taishan University, Taian, China in 2020. Currently, he is a M.S. degrees candidate in the School of Information and electronic engineering, Shandong Technology and Business University, Yantai, China. His research interests include image deblurring, image processing, and deep learning.

Zhen Hua received the B.S. and M.S. degrees in electrical automation from Taiyuan University of Technology, Taiyuan, China, in 1989 and 1992, respectively, the Ph.D. degree in electronic information engineering from China University of Mining and Technology, Beijing, China, in 2008. She is currently a professor at Shandong Technology and Business University. Her research interests include computer aided geometric design, information visualization, virtual reality and image processing.

Jinjiang Li received the B.S. and M.S. degrees in computer science from Taiyuan University of Technology, Taiyuan, China, in 2001 and 2004, respectively, the Ph.D. degree in computer science from Shandong University, Jinan, China, in 2010. From 2004 to 2006, he was an assistant research fellow at the institute of computer science and technology of Peking University, Beijing, China. From 2012 to 2014, he was a Post-Doctoral Fellow at Tsinghua University, Beijing, China. He is currently a Professor at the school of computer science and technology, Shandong Technology and Business University. His research interests include image processing, computer graphics, computer vision and machine learning.