IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 24 December 2023, accepted 13 January 2024, date of publication 16 January 2024, date of current version 26 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3354977

RESEARCH ARTICLE

Fourier Transform-Based U-Shaped Network for Single Image Motion Deblrring

JIANXIN FENG[1,2], ENGUANG HAO[1,2], YUE DU[1,2], JIANHAO ZHANG[1,2], YUANMING DING[1,2], AND HUI FANG[3], (Member, IEEE)

[1]Communication and Network Key Laboratory, Dalian University, Dalian 116622, China
[2]School of Information Engineering, Dalian University, Dalian 116622, China
[3]Department of Computer Science, Loughborough University, LE11 3TU Loughborough, U.K.

Corresponding author: Jianxin Feng (fengjianxin863@163.com)

This work was supported in part by the Interdisciplinary Project of Dalian University under Grant DLUXK-2023-ZD-001.

ABSTRACT Image deblurring is one of the fundamental tasks in image processing tasks, which can provide the necessary support for advanced tasks such as image recognition. In this paper, we propose a new deblurring model, named Efftformer model, which is specialized in the blurring elimination of motion blur. The model focuses on the recovery of detail information and edge information to provide more effective image information and better basic support for the realization of advanced tasks. In Efftformer model, firstly, we introduce a frequency domain based ReLU residual stream, which allows network to learn blur kernel level information for better restoration of original image. Secondly, we propose a cross-connection channel attention module (CCAM) to explore an effective fusion approach in multiple scales adaptively, which helps decoders to restore original image well by aggregating the semantic information in different scales. Considering the effectiveness of edge information in image recognition tasks, we enhanced the edge information in recovered image by performing a Sobel filter as well as an auxiliary edge loss function. We conducted experiments on different motion blur datasets and compared them with state-of-the-art algorithms. The experimental results show that Efftformer model proposed in this paper achieves comparable even superior performance to the state-of-the-art algorithms.

INDEX TERMS Edge enhancement, fast Fourier transform, image restoration, single image deblurring.

I. INTRODUCTION

Image blurring is a phenomenon in which an image becomes blurred due to the loss of features such as clarity, sharpness, and detail in the process of acquisition, transmission, and display. Common types of image blurring include motion blur, bokeh blur, and Gaussian blur. As one of them, motion blur is the blurring caused by the movement of camera or object, such as dragging or stretching of the image when photographing a fast-moving object. In advanced vision tasks such as target detection, motion blur can affect the performance and accuracy of detection greatly. Therefore, image deblurring can be used as a preprocessing task for some advanced tasks to improve the accuracy and robustness of their models.

In recent years, with the proliferation of various network architectures for deep learning, many network architectures have been applied to image deblurring tasks. GAN (Generative adversarial network) is often applied to image deblurring tasks [1], [2], [4], GAN learns the distribution of an image by introducing two neural networks, generator and discriminator, and generates an image similar to real image. The generator is responsible for converting low-quality blurred images into high-quality clear images, and the discriminator is responsible for determining whether the image generated by the generator is similar to the real image. By training the generator and the discriminator iteratively, GAN can generate high-quality clear images. However, GAN needs to train two networks, a generator and a discriminator, which

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

makes the training process more complicated, and the different degrees of blurring can also lead to the instability of GAN network training with poor qualities of generated images. RNN (Recurrent Neural Networks) is a neural network structure for sequential data processing that can also be used for image deblurring tasks. RNN can process sequential data by introducing a time step, which allows pixel-by-pixel processing of images [5], [6]. However, the property of RNN passing information from front to back leads to its inability to handle bi-directional dependencies well, and it is ineffective in capturing and retaining important information in the input sequence. Cascade networks [7], [8] can improve the quality of an image through multi-stage processing progressively and are therefore also applied to deblurring tasks. In an image deblurring task, a typical cascade network includes two or more sub-networks. A first sub-network is used to perform preliminary processing on the blurred image, such as removing noise or generating a candidate clear image. The second sub-network then uses the output of the first sub-network for more refined processing. However, the cascade network contains multiple sub-networks, each of which needs to be trained and optimized, which is relatively intensive in computation. So longer training and inference times, even overfitting problems can be caused. Recently, several deep learning-based methods [9], [13], [14], [22], [24], [31] have achieved SOTA results on deblurring tasks, and most of them can be regarded as variants of the classical U-Net. The self-encoder structure of U-Net consists of an encoder and a decoder, where the encoder is responsible for compressing the input image into a low dimensional feature vector and the decoder and the decoder reduces the low dimensional feature vector to the output image. Recently, several deep learning-based methods [9], [13], [14], [22], [24], [31] have achieved SOTA results on deblurring tasks, and most of them can be regarded as variants of the classical U-Net. The self-encoder structure of U-Net consists of an encoder and a decoder, where the encoder is responsible for compressing the input image into a low dimensional feature vector and the decoder reduces the low dimensional feature vector to the output image. U-Net also uses residual connections to cascade feature maps of the corresponding levels between encoder and decoder, which improves the feature representation and information reconstruction capability of the network. Our approach also uses the U-Net based encoder decoder network architecture.

In this paper, we pay special attention to the contributions that motion deblurring network models can bring to subsequent advanced tasks of target detection. Edge information and high frequency information are very important for target detection. Edge information is the information about the boundary of an object in an image, which contains information such as the shape and contour of the object. In target detection, edge information can be used to distinguish foreground and background of a target, thus helping the tracking algorithm to locate the target's position more accurately. High-frequency information can help algorithm to better distinguish target and other objects in the background, thus improving the accuracy of tracking. If the target is a vehicle, detailed information about the vehicle, such as details of the front and rear of the vehicle, can help the tracking algorithm to locate the vehicle well. Therefore, our model focuses on enhancing features from edge information and high frequency information.

With a self-attention mechanism to capture the dependencies in sequences of images, Transformer model can understand and reconstruct blurred images well. However, Transformer model has high computational complexity caused by a large number of parameters. In addition, Transformer model is prone to information loss and information distortion in the process of converting images into sequence data. Kong et al. [9] proposed frequency domain based self-attention solver (FSAS), which converts the use of Transformer in the spatial domain to the use of Transformer in the frequency domain. The detailed and structural information in the image is better preserved, and the computational complexity can be reduced greatly by the element-by-element product in the frequency domain instead of the matrix multiplication in the spatial domain to estimate the correlation in the elements of the sequence. FSAS introduces a gated mechanism based on the Joint Photographic Experts Group

(JPEG) compression algorithm in feed-forward networks to determine which features of low and high frequency information should be retained for better deblurring. But we find that this is not enough to focus on the high frequency information. In view of this, we add ReLU residual stream in feedforward network, which helps the model to learn the blur kernel information and pay more attention to the recovery of high-frequency information in the frequency domain. ReLU operation in the frequency domain also makes feedforward network have the new ability to learn the global information comparing with the general feedforward network.

In addition, we find that the hopping connection between the encoder and decoder plays an important role in propagating the information lost in downsampling for U-shaped networks to recover the complete fine information. However, the simple jump connection ignores the long-term dependence of different scale features and the problem of the spatial feature misalignment when the high-resolution encoder features are aggregated with the low-resolution decoder features. So, the cross-connection channel attention module(CCAM) is proposed in this paper to replace the original simple jump connection. As mentioned earlier, our deblur focuses more on edge information, and Sobel filter is introduced to enhance the edge information in encoder. The traditional deep learning based deblurring methods measure the greyscale map of the image simply in terms of pixels with mean square error (MSE) loss function, which does not satisfy our need to recover the high-frequency components of model. Then we introduced the auxiliary edge loss function to help the network to get closer to the true image of the target image in terms of the detailed information in the high-frequency part.

In summary, the main contribution of this work is fourfold.

1) A new feedforward network is proposed based on FSAS, which focuses more on the high-frequency components of the deblurring process and does not ignore the learning of global information.

2) CCAM (the cross-connection channel attention module) is proposed to replace the hopping connection between the encoder and decoder in the original U-Net network as the information exchange component at the encoder-decoder architectural level, which improves the network model to sense multi-scale information and improves the model performance.

3) A Sobel filter and the auxiliary edge loss function are introduced, with the former helping the model to enhance the edge information for better distinguishing foreground and background, and the latter directing the network to focus on recovering the high-frequency components in the deblurred image.

4) Efftformer encoder-decoder network model is proposed, and this deblurring model focuses more on high-frequency information and edge information to complete the advanced visual tasks. Efftformer model achieves better experimental results than the current deblurring model in quantitative experiments.

II. RELATED WORK

A. DEEP CNN-BASED IMAGE DEBLURRING METHODS

Originally, the blur kernel estimation method was commonly used for image deblurring tasks, and its main idea is to estimate the blur kernel of an image with a known blurred image and some a priori assumptions, and restore the blurred image to a clear image by processes such as inverse convolution. Sun et al. [10] proposed spatially varying kernels of motion blur by CNN. However, the blur kernel estimation method is not practical in real scenarios due to the complexity of blur features. Later, Mao et al. [11] proposed deblurring neural networks with multi-scale convolution instead of estimating blur kernels, which were trained on multiple scales and used Gaussian pyramids as inputs. Thus the detail information can be preserved, and the coarse even intermediate information can be utilized. Zamir et al. [12] proposed MPRNet architecture with three stages. In order to mine features fully at different stages, the first two stages learn the contextual information with an encoder-decoder network, and a supervised attention module is used to improve the quality of the feature images for the previous stage. The last stage acts directly on the original input image without downsampling to preserve the desired detail information. To address the problem of high computational cost in traditional image processing methods due to the multi-scale input images and stacking of sub-networks, Cho et al. [13] design a new Multiple-Input Multiple-Output U-shape network (MIMO-UNet). The encoder can acquire multi-scale input images, and the decoder can output multi-scale deblurred images. MIMO-UNet can realize the deblurring from coarse to fine with low computation cost. Chen et al. [14] designed a simple and effective baseline model by analyzing intra-block complexity and inter-block complexity in the image processing. In order to simplify the baseline model, it was revealed that the nonlinear activation function is not necessary for network. A network with nonlinear activation was further derived from the baseline model, and the proposed model has higher speed and accuracy than the previous network.

B. TRANSFORMER-BASED IMAGE DEBLURRING METHODS

Due to the excellent long sequence modeling capability, Transformer model has shown excellent performances in various advanced vision tasks such as image classification [15], target detection [16], [17], image style transfer [39] and semantic segmentation [18], [19]. Currently, Transformer model is also used to address challenges in image super-resolution [20], [25], image deblurring [21], [22], and image denoising [23], [24]. Multi-head self-attention mechanism in Transformer is complex with its global correlation between each pixel and each head, which leads the space and time complexity to be increased in a square order of magnitude with the increase of resolution. Nowaday, many scholars have proposed effective methods to solve the Transformer complexity problem due to Multi-head self-attention mechanism. Zamir et al. [22] proposed a multi-head transposed attention module that applies self-attention in a cross-channel fashion, linking cross-covariance relationships between cross-channel feature channels with interactions between each pixel, ensuring that global relationships between pixels are shown implicitly in the attention weight map. Tsai et al. [21] proposed a Stripformer architecture to reduce the high computational cost in the original self-attention mechanism, which was designed to achieve region-specific image deblurring by exploiting a priori knowledge in motion blurring with intra-strip and inter-strip attention mechanisms. Stripformer reduces the computational effort by using fewer tokens and parameters than the original Transformer, and requires less memory and computation cost. Wang et al. [24] proposed a UNet-based Transformer that introduces a new locally-enhanced window (LeWin) Transformer block, which performs non-overlapping self-attention based on windows instead of global self-attention. The computational complexity of high-resolution feature maps is significantly reduced while acquiring local context.

In addition to the image deblurring task, many models have shown greatly contributing impact in the video deblurring task. Zhang et al. [35] proposed a DeBLuRring Network (DBLRNet) for spatio-temporal learning.DBLRNet applies 3D convolution to the spatial and temporal domains to improve the video deblurring performance by jointly capturing the spatial and temporal information encoded in neighbouring frames. Li et al. [36] propose a novel implicit method to learn spatial correspondences between fuzzy frames in feature space. And to consider distant pixel

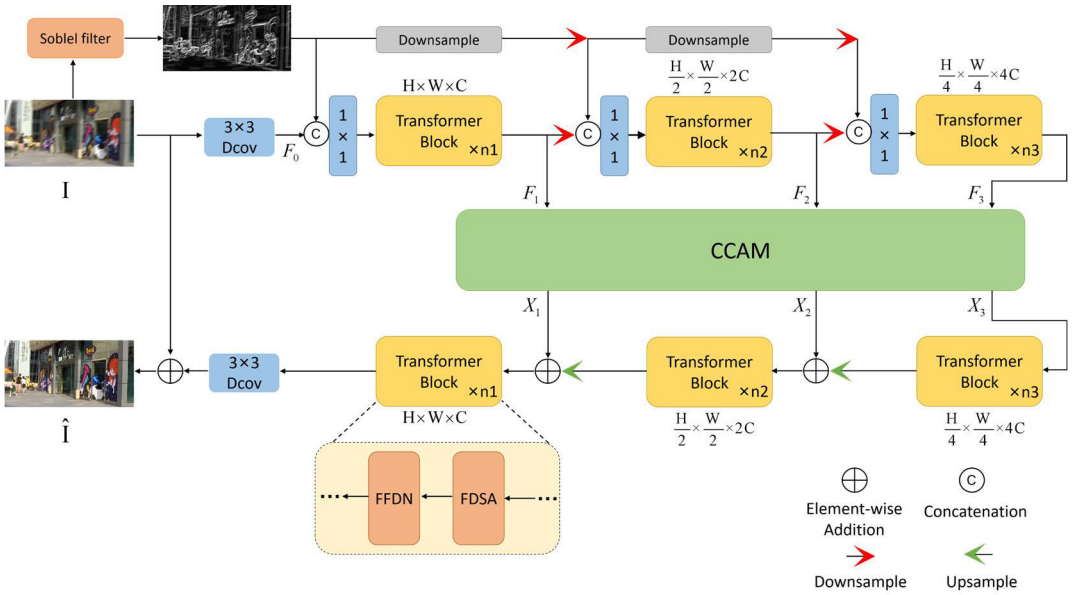

FIGURE 1. The architecture of Efftformer. Efftformer is constructed based on Unet, which is a common enconder-decoder structure. CCAM is the information exchange component between encoder and decoder. Each transformer block consists of multiple FFDN and FDSA.

displacements, they match pixel pairs in all spatial ranges between the reference frame and neighbouring frames.This confirm the benefit of modeling all-range spatial correspondence for video deblurring.

Our work uses a frequency domain -based self-attention mechanism in the encoder-decoder stage. The time and space complexity of matrix multiplication in the spatial domain is proportional to the square of the number of input sequences, i.e., it has quadratic complexity. So it means that the computation complexity will be very high when the input sequence is very large. The computation cost in the model can be reduced greatly by using the product of elements in frequency domain to replace the original matrix multiplication in spatial domain to achieve attention.

III. PROPOSED METHOD

Our objective is to present a deblurring model that focuses on details and refined edges to acquire deblurred images of superior quality. To better improve the features extracted by Transformer in the frequency-domain, we designed a new feedforward network. The new feedforward network not only takes into account the low-frequency information and the high-frequency information for recovering a clear image, but also incorporates the blur kernel information by applying ReLU operation and Fast Fourier transform in frequency-domain of the blurred image, in order to provide more effective deblurring detail information. In addition, in order to better aggregate the information between the encoder and the decoder, CCAM (cross-connection channel attention module) is designed as an adaptive information fusion component, which combines each scale feature from one encoder with other two scale features from other encoders to decrease previous semantic gaps. The fused features are subsequently added to the corresponding decoder. Moreover, the learnable Sobel-Feldman filter is integrated into the encoder part in our model to enhance the edge information in image and the auxiliary edge loss function L_{edge} is introduced to make the output image closer to the ground truth in the high-frequency detail. The model proposed in this paper is Efftformer model shown in Fig.1, where the Transformer Block consists of FFDN(feedforward network) and FDSA(frequency-domain self-attention mechanism) shown at the bottom.

Given a blurred image I, the input is first converted from an RGB image to low-level features $F_0, F_0 \in \mathbb{R}^{H\times W\times C}$ with 3×3 deep convolution, where $H \times W$ is the spatial dimension and C is the number of channels. Next, the low-level features F_0 is passed through a three-level symmetric encoder-decoder to obtain the final deep features, where the three-level encoder hierarchically reduces the size of the space while expanding the capacity of the channel with transforming F_0 into multi-scale features $F_1, F_2, F_3, F_1 \in \mathbb{R}^{H\times W\times C}$, $F_2 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 2C}$, and $F_3 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$.

In order to enhance the edge information, the blurred image I is passed through a Sobel filter to produce edge feature maps followed by a GeLU activation. After downsampling operations, the edge features with multi-scale are concatenated with F_0, F_1 and F_2 respectively. F_1, F_2 and F_3 are transformed to CCAM, which adaptively fuses the features at three different scales to obtain $X_1 \in \mathbb{R}^{H\times W\times C}, X_2 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 2C}$ and $X_3 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$. X_3 is the input features to the first level decoder. The sum of X_2 and the output features of the first level decoder is used as the input of the second level decoder. And the sum of X_1 and the output features of

the second level decoder is used as the input of the third level decoder. Finally the output features of the third level decoder are processed by 3×3 convolution, which adding with the blurred image I yields the final clear image $\hat{I}$.

A. THE FREQUENCY-DOMAIN SELF-ATTENTION MECHANISM

According to the convolution theorem, the convolution of two signals in the spatial domain is equivalent to an element-wise product of them in the frequency domain. Therefore, literature [9] considers the efficient estimation of the scaled dot-product attention by element-wise product in the frequency domain without the need to compute the matrix in the spatial domain to reduce the complexity of the model computation. Referring to [9] the FDSA (the frequency-domain self-attention mechanism) is introduced in our Transformer block shown in Fig. 2.

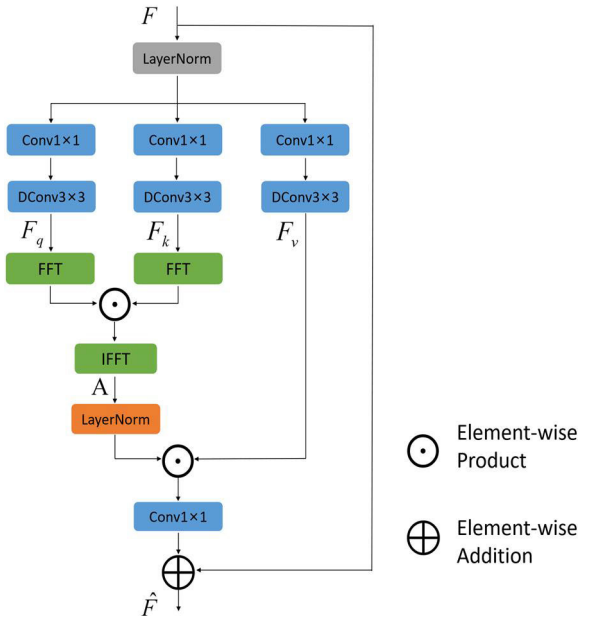

FIGURE 2. In FDSA, scaled dot-product attention is estimated by an elemental-wise product operations rather than matrix multiplication.

Assuming that given input features $F \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$, the normalization tensor $Y_F \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ is obtained after layer normalization, then the query feature F_q, the key feature F_k and the value feature F_v can be obtained after a 1×1 point-wise convolution and 3×3 depth-wise convolution.

$$F_q = W_d^Q W_p^Q Y_F \tag{1}$$

$$F_k = W_d^K W_p^K Y_F \tag{2}$$

$$F_v = W_d^V W_p^V Y_F \tag{3}$$

where $W_d^{(\cdot)}$ is a 1×1 point-wise convolution and $W_p^{(\cdot)}$ is a 3×3 depth-wise convolution. Then calculating the correlation between F_q and F_k in the frequency domain with Fast Fourier Transform using the following equation

$$A = \mathcal{F}^{-1}\left(\mathcal{F}\left(F_q\right)\overline{\mathcal{F}\left(F_k\right)}\right) \tag{4}$$

where $\mathcal{F}(\cdot)$ denotes the Fast Fourier Transform, $\mathcal{F}^{-1}(\cdot)$ denotes the Inverse Fast Fourier Transform, and $\overline{\mathcal{F}(\cdot)}$ denotes the conjugate transpose operation. Next, Attention is obtained with the following equation

$$Attention = \mathcal{L}(A) \odot F_v \tag{5}$$

where $\mathcal{L}(\cdot)$ is layer normalizing and $\odot$ denotes the element-wise product. Finally, the output of FDSA as $\hat{F}$, $\hat{F} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ can be gotten with equation (6):

$$\hat{F} = F + Conv_{1\times1}(Attention) \tag{6}$$

B. THE FEEDFORWARD NETWORK

Inspired by literature [11], the inverse Fourier transform on frequency selection (e.g., ReLU on the frequency domain) of a blurry image can act as the kernel learning pattern for the blurry image, which indicating the blur direction and blur level. In this way, the network can learn kernel-level information to perform the task of image deblurring better. As shown in Fig.3, the kernel-level information is fused with pixel-level features. The ReLU residual stream is added on the left in the feedforward network of the literature [9], and the overall structure of FFDN(the feedforward network) in the Transformer block is shown in Fig.3.

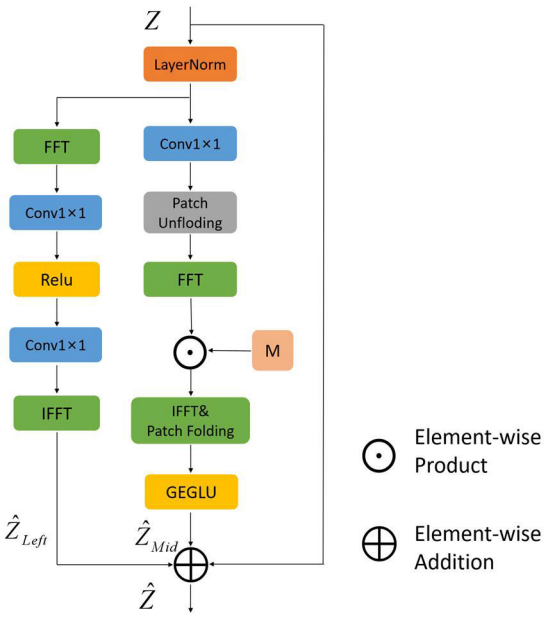

FIGURE 3. In the FFDN module, the RELU residual stream on the left are added, so that the feedforward network can fuse kernel level information, and the global context learning ability of the network can be improved by introducing nonlinear ReLU in the frequency domain.

Suppose Z, $Z \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$ is an input feature, where $\hat{H}$, $\hat{W}$, $\hat{C}$ denote the height, width and number of channels of the feature. The process of the residual stream with ReLU on the left-hand side is as follows: first, layer normalization is applied to Z, followed by applying Fast Fourier transform on the resulting tensor Y_Z. Then, two 1×1 convolutional layers are used, with one ReLU layer between them. With applying

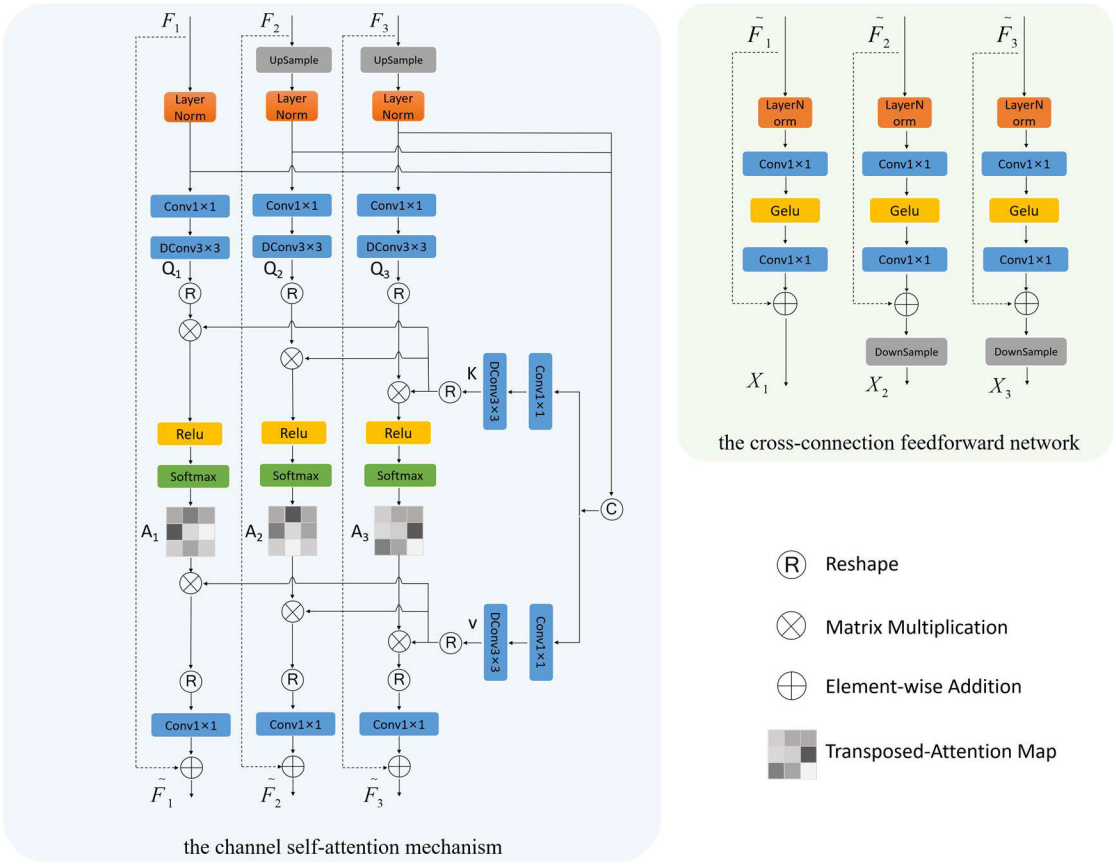

FIGURE 4. CCAM is composed of the multihead channel self-attention mechanism and feedforward network, which improves the perception ability of multi-scale information.

ReLU on frequency selection in the frequency domain and then performing an inverse Fast Fourier transform, it is possible to learn the blur kernel-level features from the blurry image directly.

$$\tilde{Z}_{Left} = ReLU(\mathcal{F}(Y_Z) \otimes W_{1\times1}^{(1)}) \otimes W_{1\times1}^{(2)} \tag{7}$$

where $\mathcal{F}(\cdot)$ denotes Fast Fourier transform, $\otimes$ denotes the convolution operation, $W_{1\times1}^{(1)}$ and $W_{1\times1}^{(2)}$ denote 1×1 convolution matrices consisting of parameters represented as complex values. Then, the inverse Fast Fourier transform is performed to obtain the output features $\hat{Z}_{Left}$ in the spatial domain.

$$\hat{Z}_{Left} = \mathcal{F}^{-1}\left(\tilde{Z}_{Left}\right) \tag{8}$$

where $\mathcal{F}^{-1}(\cdot)$ denotes the inverse Fast Fourier transform. Literature [9] argues that not all high and low-frequency information contributes to the recovery of a clear image, so a learnable quantization matrix M is introduced in the middle stream of the FFDN and it is learned by the inverse of the JPEG (Joint Photographic Experts Group) compression algorithm to determine which frequency information should be retained in the channel. In the training process, we first predefine a learnable quantisation matrix M. After obtaining a recovered simple sharp image, the error between the clear image and the original image is computed and the weight parameters in M are updated by the backpropagation algorithm. The M makes it possible to determine which high-frequency or low-frequency information is for the recovery of images with retaining the most important frequency components effectively. The middle stream of FFDN can be represented by the following equation

$$\tilde{Z}_{Mid}^{1} = \mathcal{F}\left(\mathcal{P}\left(Conv_{1\times1}\left(Y_Z\right)\right)\right) \tag{9}$$

$$\tilde{Z}_{Mid}^{2} = \mathcal{F}^{-1}\left(M \odot \tilde{Z}_{Mid}^{1}\right) \tag{10}$$

$$\hat{Z}_{Mid} = GEGLU\left(\mathcal{P}^{-1}\left(\tilde{Z}_{Mid}^{2}\right)\right) \tag{11}$$

where $\mathcal{L}(\cdot)$ denotes the layer normalization operation, $\mathcal{P}(\cdot)$ and $\mathcal{P}^{-1}(\cdot)$ denote the patch unfolding and folding operations in the JPEG compression method.

Finally, the ultimate output of FFDN can be obtained by utilizing the subsequent equation.

$$\hat{Z} = \hat{Z}_{Mid} + \hat{Z}_{Left} + Z \tag{12}$$

C. THE CROSS-CONNECTION CHANNEL ATTENTION

To bridge the gap between features at different scales, inspired by the literature [22], we propose the CCAM

(the cross-connection channel attention). CCAM bridges the semantic and resolution gap between the encoder different scale feature outputs and further adds the processed features of different scales to the decoders, which improves the network models ability to perceive multi-scale information and solves the problem of spatial feature mismatch between the high-resolution encoder features and the low-resolution decoder features. As shown in Fig.4, CCAM consists of two parts: the channel self-attention mechanism and the cross-connection feedforward network. After obtaining the encoder output features F_3, the three levels of encoder features are inputted into CCAM.

Given three different scales of encoder features $F_i(i = 1, 2, 3)$, we first perform an upsampling operation to map the features of different resolutions to the same resolution. Then, the tensors Y_i after layer normalization operation is obtained as $Y_\Sigma = Concat(Y_1; Y_2; Y_3)$ by concatenation operation. Similar to the transformer block, we can obtain the projection of the query feature Q, the key feature K, and the value feature V using a 1×1 convolution used to aggregate the cross-channel context at the pixel level and a 3×3 deep convolution operation used to emphasize the spatial context at the channel level, with the following formula.

$$Q_i = \dot{W}_d^{Q_i} \dot{W}_p^{Q_i} Y_i, i \in \{1, 2, 3\} \tag{13}$$

$$K = \dot{W}_d^{K} \dot{W}_p^{K} Y_\Sigma \tag{14}$$

$$V = \dot{W}_d^{V} \dot{W}_p^{V} Y_\Sigma \tag{15}$$

where $\dot{W}_d^{(\cdot)}$ is a 1×1 point-wise convolution and $\dot{W}_p^{(\cdot)}$ is a 3×3 depth-wise convolution. We reshape the projections of the query Q_i and key K and interact with the dot product to generate the transposed attention map A_i, and our attention mechanism is represented by the following equation:

$$\text{Attention}\,(Q_i, K, V) = \text{Softmax}\left(ReLU(\hat{Q}_i \cdot \hat{K}/\alpha)\right) \cdot \hat{V}, \quad i \in \{1, 2, 3\} \tag{16}$$

$$\text{head}_j^i = \text{Attention}(Q_i^j W_j^Q, K^j W_j^K, V^j W_j^V) \tag{17}$$

$$\text{MultiHead}\,(Q_i, K, V) = \text{Concat}(\text{head}_1^i, \cdots, \text{head}_j^i)\text{W}^\text{h} \tag{18}$$

where $\hat{Q}_i \in \mathbb{R}^{HW \times C}$, $\hat{K} \in \mathbb{R}^{C \times HW}$, $\hat{V} \in \mathbb{R}^{C \times HW}$ are obtained by reshaping Q_i, K and V, and α is a learnable scaling parameter used to control the size of the dot product of $\hat{Q}$ and $\hat{K}$, making it easier for the softmax function to produce smaller weight values. head_j^i denotes the j-th head of attention, W_j^Q, W_j^K, W_j^V and W^h are projection matrices. The channel self-attention mechanism applies the self-attention mechanism in the feature dimension rather than in the spatial dimension, and it obtains the attention map from the input features by computing the cross-covariance across the feature channels. We divide the different attention channels into different "heads" and learn different attention maps, the number of heads in our model is defaulted to 2.

Finally, the output of the channel self-attention mechanism is obtained after skip connection.

$$\tilde{F}_i = W_p \text{MultiHead}(Q_i, K, V) + F_i \tag{19}$$

where W_p is a 1×1 convolution. We get the outputs $\tilde{F}_1$, $\tilde{F}_2$, $\tilde{F}_3$. Next, we feed the obtained outputs into the cross-connection feedforward network. For the cross-connection feedforward network, The input feature $\tilde{F}_1$ is normalized through layer normalization to obtain tensors $\tilde{Y}_i, i \in \{1, 2, 3\}$. We use two 1×1 convolutions and a nonlinear Gelu activation function as the gating unit.

$$\tilde{X}_i = Conv_{1\times1}\left(GELU\left(Conv_{1\times1}\left(\tilde{Y}_i\right)\right)\right) + \tilde{F}_i, i \in \{1, 2, 3\} \tag{20}$$

Then, $\tilde{X}_1$ serves as the direct output X_1 of the CCAM, while $\tilde{X}_2$ and $\tilde{X}_3$ are downsampled to generate X_2 and X_3, respectively, as outputs of the CCAM module.

The lowest resolution output feature X_3, is used as the input to the first level of the decoder, and X_1 and X_2 are the two remaining outputs, which are summed with the corresponding decoder output features, respectively, and used as the input to the next level of decoder.

D. THE SOBEL FILTER

Inspired by the literature [3], to enhance the edge information of the model, we use learnable Sobel filters as our edge enhancement block. Sobel filters are often used in edge detection algorithms due to the fact that it helps to enhance the edges.

As shown in Fig.5, different from the traditional fixed-value Sobel operator, a learnable parameter α' is defined, which is called Sobel factor. The value of α' can be adaptively adjusted to extract edge information of different intensity during the optimization of training process. Besides, we define four types of filters as a group to extract edge information of different directions including vertical, horizontal, main diagonal and secondary diagonal directions.

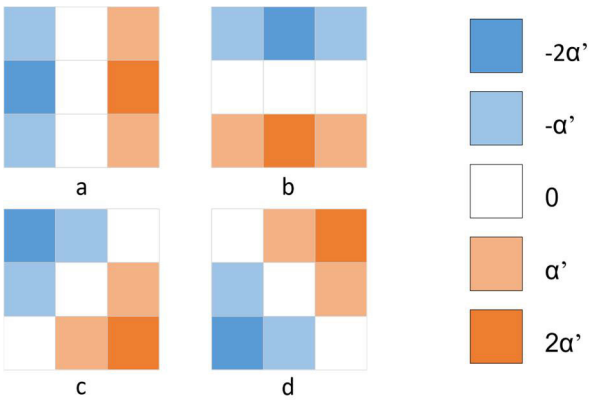

FIGURE 5. We use four different Sobel filters including: vertical (a), horizontal (b), and diagonals (c,d). α' is a learnable paramerer.

The Sobel filter is added in the model. The input image undergoes Sobel convolution for edge feature extraction. The

extracted features are then activated using the GeLU function. The input feature of the 1th-level encoder directly concats with edge features, while this of the 2th-level encoder concats with the edge features after primary downsampling. Lastly, the input feature of the 3th-level encoder concat with the edge features after secondary downsampling.

E. THE AUXILIARY EDGE LOSS FUNCTION

The auxiliary edge loss function helps the network to get closer to the target image in the high frequency part. On the same network model, it shows better de-coarsening performance than the traditional MSE loss function, and our auxiliary edge loss function uses the Sobel operator to improve the performance of the model. The Sobel operator consists of horizontal and vertical Sobel kernels. Each Sobel kernel can be decomposed as the products of an averaging and a differentiation kernel. The horizontal and vertical Sobel kernels G_x, G_y can be written as

$$G_x = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}^{\mathrm{T}} * \begin{bmatrix} +1 & 0 & -1 \end{bmatrix} \tag{21}$$

$$G_y = \begin{bmatrix} +1 & 0 & -1 \end{bmatrix}^{\mathrm{T}} * \begin{bmatrix} 1 & 2 & 1 \end{bmatrix} \tag{22}$$

Sobel operation detects clear and emphasizing edge maps, since the differential kernel $\begin{bmatrix} +1 & 0 & -1 \end{bmatrix}$ obtains the gradient magnitude of the image and the averaging kernel $\begin{bmatrix} 1 & 2 & 1 \end{bmatrix}$ reduces unnecessary noise in the image. The high-frequency components of the image which are acquired by Sobel operation as follows:

$$f_{sobel}(I) = \sqrt{(G_x * I)^2 + \left(G_y * I\right)^2} \tag{23}$$

Typically, the Sobel edge mapping of a sharpened image contains sharpened edges, while the blurred image contains no sharpened edges. Since the goal of deblurring is to restore the sharp image, we directly minimize the distance between the resultant output image and the true value image in the Sobel edge space. Therefore, the proposed auxiliary edge loss function can be expressed as follows:

$$L_{edge} = \frac{1}{N} \left\| f_{sobel}(\hat{I}) - f_{sobel}(I_{\text{truth}}) \right\|_1 \tag{24}$$

where $\hat{I}$ and I_{truth} denote the resultant output image and ground truth image, the loss values are normalized by the total number of pixel elements N. The loss values are averaged over each pixel to compare the magnitude of the loss values for images of different resolutions.

IV. EXPERIMENTS

A. DATASETS

We evaluate three commonly used motion blur datasets, including the GoPro dataset [26], the HIDE dataset [27], the RealBlur dataset [28], the MC-Blur dataset [29] and the RWBI dataset [30]. We follow the protocol of existing methods for fair comparison.

B. LOSS FUNCTION

As with other multiscale deblurring networks, we use a multiscale content loss function, and the coarse-to-fine approach essentially requires that each intermediate output become a clear image of the corresponding scale. Therefore, we train our network so that the intermediate outputs should form a pyramid of clear images. The L1 loss is applied to each level of the pyramid. Thus, the content loss function is defined as follows.

$$L_{cont} = \sum_{k=1}^{K} \frac{1}{t_k} \left\| \hat{S}_k - S_k \right\|_1 \tag{25}$$

where K represents the number of levels. $\hat{S}_k$, S_k denote the model output and ground truth image at scale level k, respectively. The loss at each scale is normalized by the number of elements t_k.

In a recent study [13], auxiliary loss functions other than the content loss have also been proposed to improve the performance, since the aim of deblurring is to recover the lost high frequency components, it is crucial to reduce the difference in the frequency space, so we introduce the Multi-scale Frequency Reconstruction (MSFR) loss function, which is defined as follows:

$$L_{MSFR} = \sum_{k=1}^{K} \frac{1}{t_k} \left\| \mathcal{F}(\hat{S}_k) - \mathcal{F}(S_k) \right\|_1 \tag{26}$$

where $\mathcal{F}$ stands for the Fast Fourier transform that transforms the image signal into a frequency domain signal.

the auxiliary edge loss function as

$$L_{edge} = \frac{1}{N} \left\| f_{sobel}(\hat{I}) - f_{sobel}(I_{\text{target}}) \right\|_1 \tag{24}$$

The final loss function for training our network is defined as follows:

$$L_{total} = L_{cont} + \lambda_1 L_{MSFR} + \lambda_2 L_{edge} \tag{27}$$

where the weights of the auxiliary loss components, were set to 0.1 and 0.05 in this experiment, respectively.

C. PARAMAETER SETTINGS

Our model uses a three-level encoder-decoder structure with [6,6,12] transformer blocks and [48,96,192] channels from level 1 to level 3. We set the initial learning rate to 10^{-3} and update it after 6×10^6 iterations with a cosine annealing strategy [37], which steadily decreases the learning rate to 10^{-7}. During training, we use horizontal and vertical flipping to increase the data. The default patch size is set to 256×256 pixels and the batch size is set to 16. In addition, the quantization matrix M of the feedforward network and the patch size of the self-attention mechanism in the Transformer block are both set to 8×8. During training, we use the Adam optimizer [38] as our stochastic gradient descent algorithm, and the quantization matrix M is minimized by jointly learning it along with other parameters to minimize the loss function.

TABLE 1. Performance comparisons in GoPro dataset.

Methods	PSNRs	SSIMs	Parameters(M)
Restormer[22]	32.92	0.961	26.1
Uformer[24]	33.06	0.967	50.9
Stripformer[21]	33.08	0.962	19.7
NAFNet[14]	33.71	0.966	67.9
MIMO-Unet+[13]	32.45	0.956	16.1
MAXIM[34]	32.86	0.961	22.2
GRL-B[32]	33.93	0.968	20.2
Ours	34.26	0.963	25.7

TABLE 2. Performance comparisons in HIDE dataset.

Methods	PSNRs	SSIMs	Parameters(M)
Restormer[22]	31.22	0.943	26.1
Uformer[24]	30.83	0.952	50.9
Stripformer[21]	31.03	0.935	19.7
NAFNet[14]	31.31	0.942	67.9
FFTformer[9]	31.62	0.945	16.6
MIMO-Unet+[13]	29.99	0.930	16.1
GRL-B[32]	31.65	0.947	20.2
Ours	31.73	0.943	25.7

TABLE 3. Performance comparisons in RealBlur dataset.

Methods	Realblur-R		Realblur-J	
	PSNRs	SSIMs	PSNRs	SSIMs
MPRNet[12]	39.31	0.972	31.76	0.922
MAXIM[34]	39.45	0.962	32.84	0.935
DeepRFT+[33]	39.84	0.972	28.97	0.884
Restormer[22]	36.19	0.957	28.96	0.879
FNAFNet[11]	36.07	0.955	28.78	0.879
Uformer[24]	36.22	0.957	29.06	0.884
Ours	40.03	0.973	32.53	0.933

D. EVALUATION INDICATORS

We compare the present method with existing methods and evaluate the quality of the recovered images using PSNR and SSIM, PSNR (Peak Signal-to-Noise Ratio) denotes the peak signal-to-noise ratio, which is a common measure of image reconstruction quality. It calculates the logarithmic inverse of the mean square error (MSE) between the original and reconstructed images to measure the relative difference in image quality. The higher the value of PSNR, the lower the difference between the reconstructed image and the original image, and the higher the quality of the reconstructed image. SSIM (Structural Similarity Index) is another commonly used image quality evaluation index, which not only considers the mean square error of the image, but also the structural similarity of the image. SSIM evaluates the image quality by calculating the similarity in three aspects: brightness, contrast and structure. The value of SSIM ranges from 0 to 1, and the smaller the difference from 1, indicates that the more similar the structure and content of the two images, the higher the quality of the reconstructed image.

E. EVALUATION ON DATASETS

1) EVALUATED ON GOPRO DATASET

Firstly our method is evaluated on GoPro dataset [26]. For fair comparison, we follow the protocol of this GoPro dataset, and the quantitative evaluation results are shown in Table 1. Our method generates the results with the highest PSNR and SSIM values. Compared to NAFNet in the CNN method, PSNR in our method is improved by 0.55dB.

Fig.6 illustrates the comparisons on GoPro dataset for vision. As previously stated, our method places greater emphasis on details and edges. The results show that our approach performs better than other methods in recovering license plate and chair details obviously.

Fig.6 illustrates the comparisons on GoPro dataset for vision. As previously stated, our method places greater emphasis on details and edges. The results show that our approach performs better than other methods in recovering license plate and chair details obviously.

2) EVALUATED ON HIDE DATASET

To demonstrate the impressive generalization capability of our technique, we assessed our model on the motion-blur HIDE dataset [27], which primarily features humans. Compared to NAFNet in the CNN method, PSNR in our method is improved by 0.34dB.

Fig.7 presents visual comparisons on HIDE dataset. Our method outperforms the others in recovering zippers and clothing on the picture characters and produces clearer images of the characters' feet compared to the other methods. This visually demonstrates the effectiveness of our method in focusing on detailed information.

3) EVALUATED ON REALBLUR DATASET

In addition, we evaluated our method on RealBlur dataset [28], which consists of RealBlur-R and RealBlur-J test sets, both comprising image blur samples from real scenes. Table 3 shows that our method achieves higher PSNR values than other deblurring methods. Compared with the latest deblurring method DeepRFT+ [11], our approach achieves PSNR is 0.19 dB higher on Realblur-R testset. Fig. 8 illustrates a visual comparison of various deblurring methods, indicating that our approach restores blurred number plates in low-light settings satisfactorily.

4) EVALUATED ON OTHER DATASETS

To further demonstrate the generalisation performance of our model, we evaluated our model on the MC-Blur dataset [29]. The MC-Blur dataset contains four subsets which are RHM, UHDM, LSD and RMBQ. Specifically, RHM subset provides motion blurred images synthesised in variable fps video frames. UHDM subset is an Ultra-High-Definition (UHD)

FIGURE 6. Visual comparisons of GoPro dataset [26], our method has better deblurring effect for detail information of the objects.

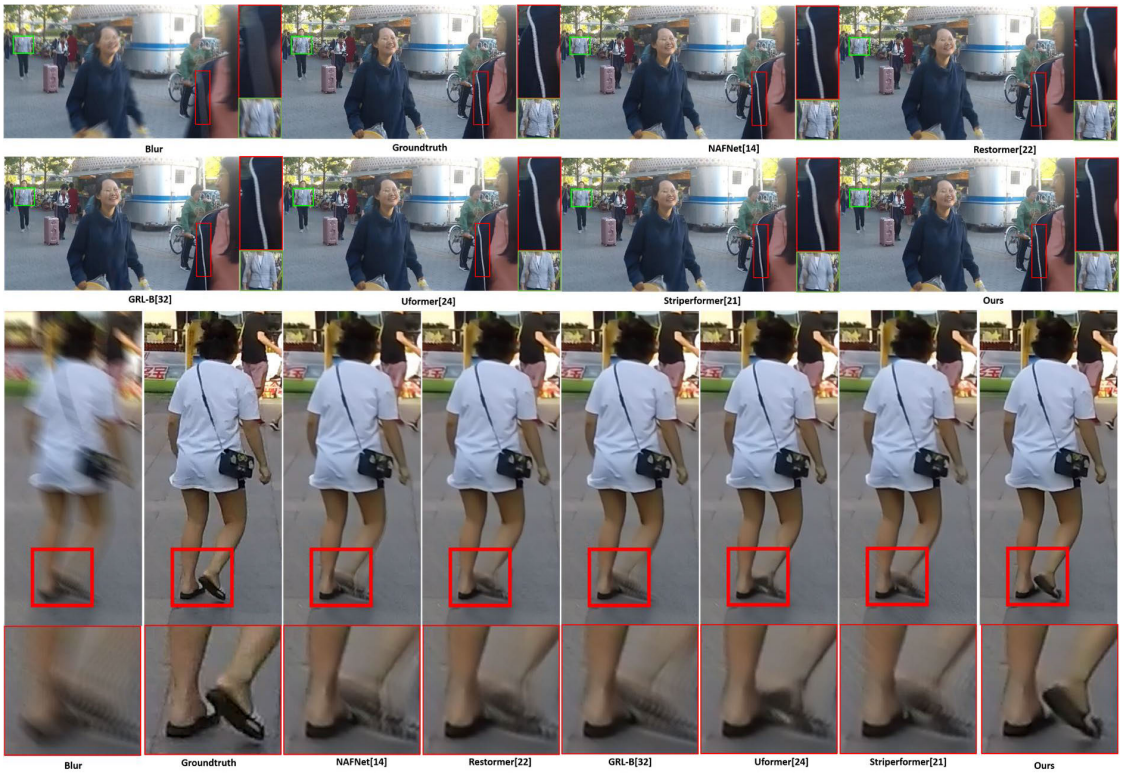

FIGURE 7. Visual comparisons of HIDE dataset [27], our method has better deblurring effect for detail information of the characters.

image deblurring dataset, which blurring kernels of various sizes are used to convolve with clear images to obtain blurred images. LSD subset is a defocus blurry dataset. RMBQ subset provides large-scale, real blurry images.

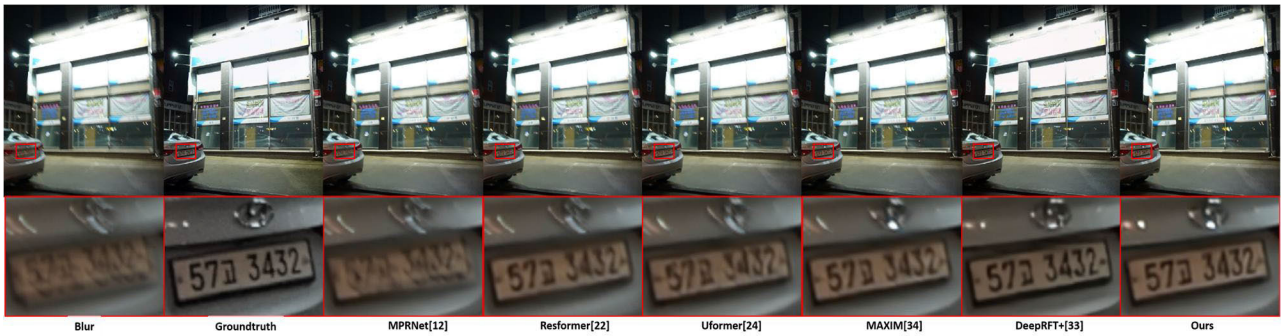

FIGURE 8. Visual comparisons of RealBlur dataset [28], our method has better deblurring effect for detail information in low light environment.

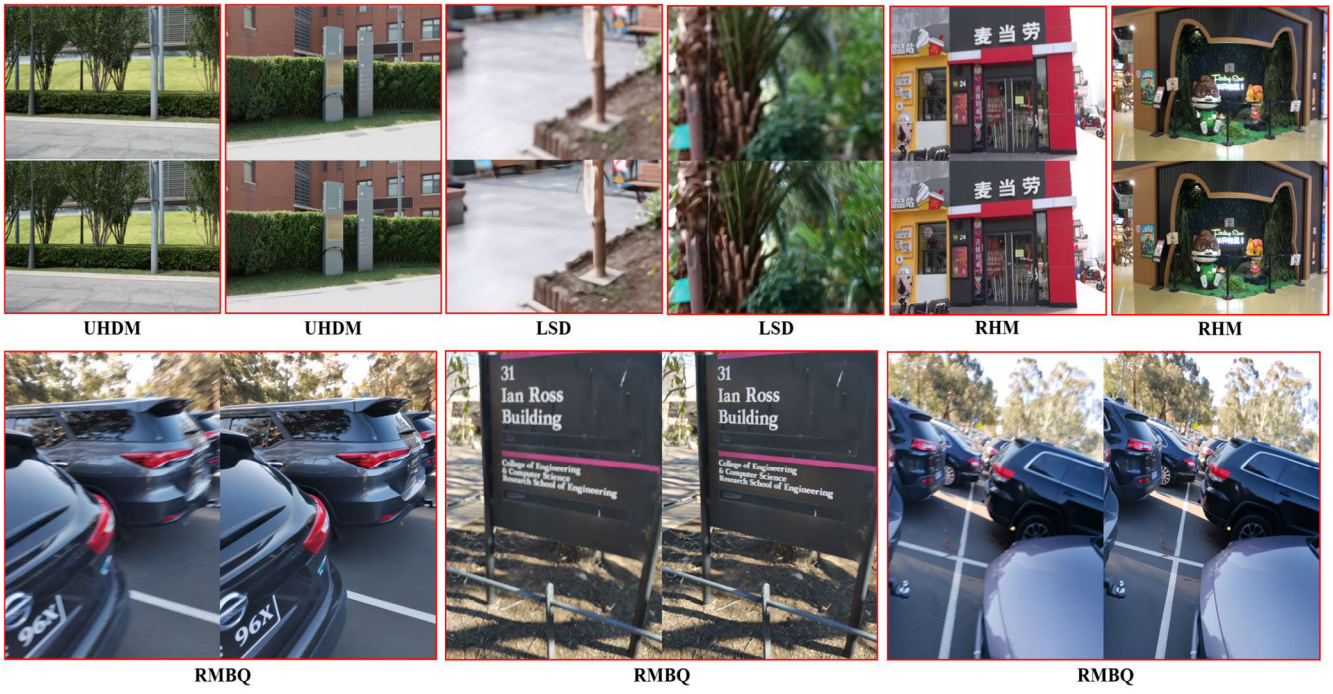

FIGURE 9. Visual comparisons of MC-Blur dataset [29]. MC-Blur dataset consists of four sub-datasets engineered with different reasons for blurring UHDM, LSD, RHM and RMBQ respectively. In the image pairs of UHDM, LSD and RHM, the upper images are the blurry images and the lower images are the recovery images.In the image pairs of RMBQ, the left images are the blurry images and the right images are the recovery images.

FIGURE 10. Visual comparisons of RWBI dataset [30]. In the image pairs, the upper images are the blurry images and the lower images are the recovery images.

Shown in Fig.9, Efftformer can not get good deblurring results for defocus blurry in LSD subset due to the different causes and features of out-of-focus blur and motion blur. However, Efftformer can observe

FIGURE 11. Visual comparisons of blurry licence plate images. In the image pairs, the left images are the blurry images and the right images are the recovery images.

obvious deblurring results in the real blurring RMBQ subset.

In addition, evaluation experiments are made in RWBI dataset [30] with a Real-World Blurry Images captured with different hand-held devices. Efftformer has an excellent recovery effect for real world blurry images shown in Fig.10.

5) POTENTIAL REAL-WORLD APPLICATIONS

In this section we try to explore the practical applications of Efftformer.Efftformer can be applied in many real world applications as an image deblurring model. In photography and image processing, Efftformer can be used to repair blurry images to improve the quality and clarity of photos caused by camera shaking, object motion, or inaccurate focusing. In blurry document processing, Efftformer can restore document images with clarity and readability.

We discuss the specific application of Efftformer in removing blur of car licence plate. Licence plate is a unique and accurate identification of a car, but blurry licence plate images are not conducive to licence plate information recognition, segmentation and re-identification of the same licence plate. CCPD dataset [40] is selected to restore the clarity of the blurred licence plate information using Efftformer in th real world. Shown in Fig.11, Efftformer has an excellent recovery effect for deblurring licence plate information. This can be an effective contribution to traffic management.

F. ABLATION STUDY

Our ablation studies are designed to prove the effectiveness of our added components. Specifically, we trained our method and all the baselines on GoPro dataset.

1) First, we prove the validity of ReLU residual stream in the FFDN. We compared different FFDN with and without residual stream, as shown in Fig. 12 respectively. a is feedforward network without ReLU residual stream, b is ReLU residual stream only, and c is our method that feedforward network with ReLU residual stream.

The ReLU residual stream only approach results in a degradation of model performance, with a 1.51 dB reduction in

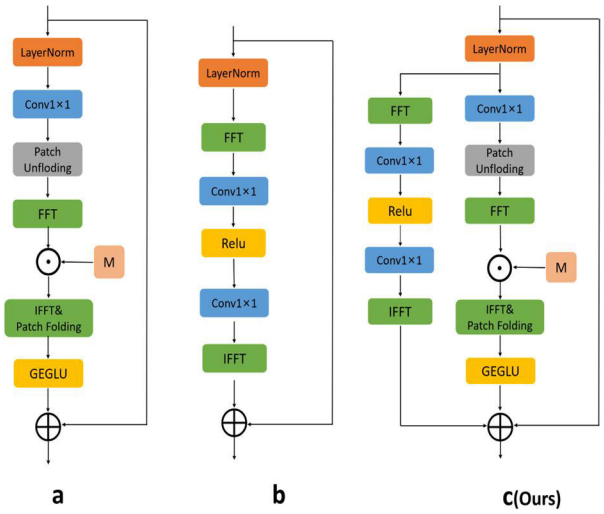

FIGURE 12. Ablation experiments of ReLU residual stream in FFDN. a is without ReLU residual stream, b is ReLUc stream only, and c is our proposed method.

TABLE 4. Ablation study results.

	PSNRs	SSIMs	Parameters(M)
w/o ReLU residual stream	34.07	0.962	23.3
Only ReLU residual stream	32.75	0.946	21.9
w/o CCAM	34.13	0.962	25.1
SD	33.98	0.961	56.0
Ours	34.26	0.963	25.7

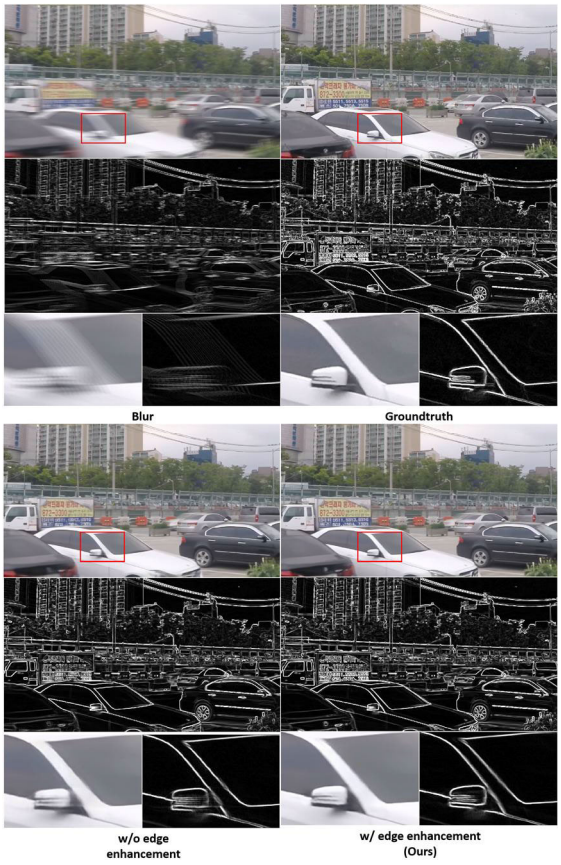

FIGURE 13. Visual comparison of edge enhancement with the Sobel filter and the auxiliary edge loss function.

PSNR values (34.26 vs. 32.75). It shows that ReLU residual stream need to be trained in conjunction with pixel-level methods, or it does not result in effective recovery because of focusing on kernel-level information only. PSNR of our method is improved by 0.19 dB (34.26 vs. 34.07) compared to the method without ReLU residual stream shown in Table 4.

2) In order to demonstrate the effectiveness of CCAM, we replaced CCAM with a general skip connection (i.e., without CCAM) and compared the performance gap. Our method improved PSNR by 0.13 dB compared to the general skip connection (34.26 vs. 34.13) and it only increases the number of parameters by 0.6M. shown in Table 4. Compared to other components, CCAM achieves a significant performance improvement with less memory.

3) In order to test the efficiency of the scaled dot-product attention across three-level encoder-decoder in the frequency domain, our method is compared with the baseline method in the spatial domain (SD for short). Specifically, Swin attention in Swin Transformer [15] is choose to replace the frequency domain attention in our three-level encoder-decoder. The quantitative evaluation in GoPro dataset shown in Table 4 is that the application of scaled dot-product attention in SD reduces PSNR by 0.28 dB (34.26 vs. 33.98).

4) With the visual comparison in Fig.13, the Sobel filter as well as the edge loss function for edge enhancement are compared. By visual comparison, it can be observed that with the addition of the Sobel filter and the edge loss function, a much clearer edge structure of the car can be obtained to recover the image closer to the target image in some extent.

V. CONCLUSION

In this paper, we propose a new motion deblurring framework, Efftformer, which pays more attention to the detailed information and edge information of blurred images. ReLU residual stream are introduced in the encoder-decoder section to enable the model to learn kernel-level information, and by focusing on the kernel-level information versus the pixel-level information, better de-blurred images are obtained. In addition, CCAM adaptively blends the semantic and resolution gaps between different scales, more fully utilizing both low and high level information to obtain clear images with better results. In order to handle the edge information, the Sobel filter is introduced at the decoder stage for edge enhancement, where the edge loss function utilizes the high frequency components in the image to make the blurred image closer to the target image. Extensive experiments have shown that Efftformer achieves state-of-the-art performance with high validity and generalizability in six public motion blur datasets. In addition, Efftformer has less paramaters

under the same performance compaired with other deblurring models. In the current work, Efftformer is specifically designed for motion-blurred images, and in the next work, we will further explore the potential of our model for other blur types and other image restoration tasks, such as improving performance in Efftformer for defocus blur in doucument images.

REFERENCES

[1] O. Kupyn, V. Budzan, M. Mykhailych, 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.

[2] 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. 8877–8886.

[3] T. Liang, Y. Jin, Y. Li, and T. Wang, "EDCNN: Edge enhancement-based densely connected network with compound loss for low-dose CT denoising," in *Proc. 15th IEEE Int. Conf. Signal Process. (ICSP)*, vol. 1, Dec. 2020, pp. 193–198.

[4] J. Feng and Y. Han, "An underwater image enhancement strategy based on pyramid attention mechanism," in *Proc. 3rd Int. Conf. Opt. Image Process. (ICOIP)*, Bellingham, WA, USA, Aug. 2023, pp. 334–341.

[5] 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. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 327–343.

[6] J. Zhang, J. Pan, J. Ren, Y. Song, L. Bao, R. W. H. Lau, and M.-H. Yang, "Dynamic scene deblurring using spatially variant recurrent neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2521–2529.

[7] 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, Jul. 2016.

[8] J. Pan, H. Bai, and J. Tang, "Cascaded deep video deblurring using temporal sharpness prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3040–3048.

[9] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5886–5895.

[10] 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.

[11] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, "Intriguing findings of frequency selection for image deblurring," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, no. 2, pp. 1905–1913.

[12] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14816–14826.

[13] 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. (ICCV)*, Oct. 2021, pp. 4621–4630.

[14] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 17–33.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[17] J. Zhang, J. Xie, N. Barnes, and P. Li, "Learning generative vision transformer with energy-based latent space for saliency prediction," in *Proc. NIPS*, vol. 34, 2021, pp. 15448–15463.

[18] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, and D. Xu, "Multi-class token transformer for weakly supervised semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4300–4309.

[19] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[20] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[21] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 146–162.

[22] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[23] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12294–12305.

[24] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.

[25] J. Wang, B. Wang, X. Wang, Y. Zhao, and T. Long, "Hybrid attention-based U-shaped network for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

[26] 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. 257–265.

[27] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao, "Human-aware motion deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5571–5580.

[28] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *Computer Vision—ECCV 2020*. Cham, Switzerland: Springer, Aug. 2020, pp. 184–201.

[29] K. Zhang, T. Wang, W. Luo, W. Ren, B. Stenger, W. Liu, H. Li, and M.-H. Yang, "MC-blur: A comprehensive benchmark for image deblurring," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[30] K. Zhang, W. Luo, Y. Zhong, L. Ma, B. Stenger, W. Liu, and H. Li, "Deblurring by realistic blurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2734–2743.

[31] C. Williams, F. Falck, G. Deligiannidis, C. Holmes, A. Doucet, and S. Syed, "A unified framework for U-Net design and analysis," 2023, *arXiv:2305.19638*.

[32] Y. Li, Y. Fan, X. Xiang, D. Demandolx, R. Ranjan, R. Timofte, and L. Van Gool, "Efficient and explicit modelling of image hierarchies for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18278–18289.

[33] X. Mao, Y. Liu, W. Shen, Q. Li, and Y. Wang, "Deep residual Fourier transformation for single image deblurring," vol. 2, no. 3, p. 5, 2021, *arXiv:2111.11745*.

[34] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MAXIM: Multi-axis MLP for image processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5759–5770.

[35] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial spatio-temporal learning for video deblurring," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 291–301, Jan. 2019.

[36] D. Li, C. Xu, K. Zhang, X. Yu, Y. Zhong, W. Ren, H. Suominen, and H. Li, "ARVo: Learning all-range volumetric correspondence for video deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7717–7727.

[37] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[39] J. Feng, G. Zhang, X. Li, Y. Ding, Z. Liu, C. Pan, S. Deng, and H. Fang, "A compositional transformer based autoencoder for image style transfer," *Electronics*, vol. 12, no. 5, p. 1184, Mar. 2023.

[40] Z. Xu, W. Yang, A. Meng, N. Lu, H. Huang, C. Ying, and L. Huang, "Towards end-to-end license plate detection and recognition: A large dataset and baseline," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 255–271.

JIANXIN FENG received the Ph.D. degree from Northeastern University, China, in 2005. From 1999 to 2012, she was a Teacher with the Information Science and Engineering Institution, Northeastern University. From 2018 to 2019, she was a Visiting Scholar with the Department of Computer Science, Liverpool John Moores University. She is currently an Associate Professor with the Information Engineering College, Dalian University, China. Her current research interests include network protocol, wireless communication, and network optimization.

ENGUANG HAO received the B.S. degree in network engineering from Liaocheng University, Liaocheng, China, in 2021. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interests include image restoration and single-object tracking.

YUE DU received the B.S. degree in electronic information engineering from Dalian University, Dalian, China, in 2022, where she is currently pursuing the master's degree with the Communication and Network Laboratory. Her current research interests include UAV communication technology and image processing.

JIANHAO ZHANG received the B.S. degree in computer science and technology from Shenyang Normal University, Shenyang, China, in 2021. He is currently pursuing the master's degree with the Communication and Network Laboratory, Dalian University. His current research interests include time series forecasting and image processing.

YUANMING DING received the Ph.D. degree from Keio University, Japan, in 2004. From November 2004 to November 2016, he was a Postdoctoral Fellow with JSPS. Since 2009, he has been a Professor with the Information Engineering College, Dalian University, China. His research interests include satellite network communication, underwater optical communication, and network technologies.

HUI FANG (Member, IEEE) received the B.S. degree from the University of Science and Technology, Beijing, China, in 2000, and the Ph.D. degree from the University of Bradford, U.K., in 2006. He is currently with the Department of Computer Science, Loughborough University. Before, he carried out research at several world-leading universities, such as the University of Oxford and Swansea University. His research interests include computer vision, image/video processing, pattern recognition, machine learning, data mining, scientific visualization, visual analytics, and artificial intelligence.