[1]Fukang Tong
[2,*]Jia He
[3]Jin Jin

Image Deblurring Based on Efficient Transformer and Multi-scale CNN

Journal of Electrical Systems

Abstract: - The study of image deblurring techniques in dynamic scenes is a high-profile research direction. Recently, given the excellence of Convolutional Neural Networks (CNNs) in extracting feature information, it has become a common and effective practice to utilize them for blurred image restoration work. However, CNN can only model local information and has a limited receptive field, which inhibits the deblurring effect. Transformer can model global information, so it can be combined with CNN to expand the receptive field and enhance the deblurring effect. Unfortunately, as the spatial resolution of the input image increases, the computational complexity of the Transformer increases dramatically, showing a trend of square-level growth, which makes it difficult to cope with the task of processing high-resolution images. To address the above problems, this paper proposes an image deblurring network based on efficient Transformer and multi-scale CNN called ET-MIMO-UNet. The local spatial features are extracted using multi-scale CNN and embedded into the global characteristics of the Transformer, modelling both local and global information. To solve the problem of difficult training due to large image size and to improve the computational efficiency, an efficient Transformer layer (ETL) is designed, which contains a multi-dconv head transposed attention (MHTA) and a gated-dconv feed-forward network(GFFN). In addition, a multi-layer feature fusion block (MFFB) is introduced to fuse full-scale features and reduce feature loss. On the GoPro test dataset, compared with the MIMO-UNet base network, the PSNR of the three models of ET-MIMO-UNet is improved by 0.39dB, 0.54dB, and 0.66dB, respectively; ET-MIMO-UNet reduces the number of parameters by half compared with MPRNet. The experimental data fully proved that the method demonstrated a significant processing effect in coping with the image-blurring problem in dynamic scenes.

Keywords: Image Deblurring, Multi-scale CNN, Transformer, Self-attention.

I. INTRODUCTION

When taking pictures, the quality of the image is often degraded by shaky camera equipment or the rapid movement of objects, resulting in blurred images. Image deblurring uses image processing techniques to restore a blurred image to an image with an explicit edge structure and rich details. Due to the many causes of image blurring, the image deblurring problem is a highly ill-posed problem with numerous unknown solutions[1].

In the image deblurring tasks, most traditional methods[2-4] solve the problem by modelling the deblurring problem as estimating the blur kernel. However, in the real-world scene, the blur kernel is unknown and very complex. Therefore, the estimation of the blur kernel is very tricky, and ultimately, the recovery of the blurred image is poor due to the inaccurate estimation of the blurring kernel.

In recent years, image deblurring methods[5,6] based on CNN have been widely studied. Earlier, researchers used CNNs to estimate the blurring kernel of an image and devised a two-step deblurring method: first, the blurring kernel is estimated using CNNs; then, a deconvolution operation is performed using this kernel to remove image blur [7,8]. However, they depend more on the blur kernel and find it difficult to cope with various blur types. In contrast, today's CNN-based image deblurring methods are more advanced, and they directly learn the complex mapping relationship between blurred and clear images in an end-to-end manner[9-13], thus handling various blurring situations more flexibly.

Although CNN has achieved high performance in single-image deblurring, it also brings two significant problems: (1) the convolutional operator has a limited receptive field, which makes it challenging to capture the information of remote pixels; (2) convolutional kernels have static weights in the inference, which are not able to adapt to the input flexibly. Different from the convolution operation that models the local feature, Transformers can dynamically model the global contexts by computing the correlations of one token to all other tokens[14]. So, the natural idea is to combine the Transformer with CNNs to expand the receptive field, modelling the image's local and global information to enhance the deblurring effect[15].

Based on the above objectives, this paper proposes an image deblurring network based on an efficient Transformer and Multi-scale CNN named ET-MIMO-UNet. We apply an encoder-decoder structure to extract

[1] College of Computer Science, Chengdu University of Information Technology, Chengdu, Sichuan 610225, China
[2] College of Computer Science, Chengdu University of Information Technology, Chengdu, Sichuan 610225, China
[3] College of Computer Science, Chengdu University of Information Technology, Chengdu, Sichuan 610225, China
*Corresponding author: Jia He

multi-scale features as most deblurring models do and use a coarse-to-fine approach to recover the image. Multi-scale CNNs encode local features and then cascade with the Transformer, allowing the network to encode the input image's global and local information. An efficient transformer layer (ETL) is introduced to improve computational efficiency and solve the problem of input features that are too large of a scale and are difficult to train. The efficient Transformer consists of a multi-dconv head transposed attention (MHTA) and gated-dconv feed-forward network (GFFN). MHTA uses channel-based self-attention to reduce its computational complexity to linear. GFFN controls the information flow and extracts critical detailed information. A multi-layer feature (MLFF) fusion block is introduced to fuse full-scale information to reduce feature loss.

The following are the main points of contribution to this paper:

1. Multi-scale CNN cascaded Transformer: This combination helps preserve the local details of the image and explore global features over long distances with significant advantages.
2. An efficient Transformer layer consists of MHTA and GFFN. MHTA reduces the computational complexity to linear by using a self-attentive mechanism based on channels rather than spatial dimension; GFFN suppresses invalid information and retains critical information.
3. A multilayer feature fusion block is used to fuse full-scale features and reduce feature loss.
4. We propose an image deblurring network based on efficient Transformer and multi-scale CNN, named ET-MIMO-U-Net, and verify the effectiveness of the model through extensive experiments.

II. Related Work

A. Deep Learning for Image Deblurring

Recently, deep learning has become a significant approach in the field of image deblurring. Sun et al.[7] proposed using CNN to estimate the spatial variation kernel of motion blur to remove non-uniform blur. Still, due to the complexity of the blur characteristics, the blur kernel estimation method cannot recover the blurred image well in the actual scene. Kupyn et al.[9] proposed DeblurGAN to achieve deblurring in a single scale based on generative adversarial network (GAN) and content loss. As a pioneering work, Nah et al.[10] introduced DeepDeblur, a deep multi-scale CNN network for de-blurring dynamic scenes based on the coarse-to-fine strategy, which extracts the multiscale information of an image without estimating any blurring kernel and recovers a clear image from a blurred image directly. However, this approach's computational time is long because the design does not share parameters across multiple scales. To solve this problem, encoder-decoder structures with skip connections are introduced to share parameters and capture contexts, e.g., PSS-NSC[16], MT-RNN[17], and MIMO-Uet[18]. Zhao et al.[19] proposed a lightweight and real-time unsupervised image blind deblurring method, FCL-GAN, with no image domain or resolution restriction, which guarantees lightweight and performance advantages. Tsai et al.[20] proposed the BANet model by employing a multicore strip-pooling attention structure to extract multiscale features. Although these models have achieved satisfactory results, they mainly focus on local features while ignoring global features, which limits the deblurring effect of the models to some extent.

B. Vision Transformer

Proposed initially for Natural Language Processing (NLP)[21,22], The Transformer employs multi-head self-attention to capture global relationships between individual tokens. Due to its powerful global modelling capabilities, Transformer is used in a variety of vision applications such as image classification[23], object detection[24], semantic segmentation[25], inpainting[26], and super-segmentation[27]. Let's consider vision Transformers (ViT)[23] as an image classification example. ViT treats images similarly to language sequences by introducing the concept of patches. The input image is segmented into patches, and the transformer structure is subsequently employed to capture the inter-patch relationships. Nevertheless, disregarding convolution entirely is not advisable as the Transformer solely relies on global self-attention, thereby overlooking the capture of finer, local details. As stated by Zhang et al.[11] global and local information is crucial for deblurring real images. Therefore, this paper introduces a Transformer into CNN networks to model global and local detail information.

Unfortunately, it is not feasible to introduce the transformer directly into the CNN network because the computational complexity of self-attention (SA) in the transformer increases exponentially as the number of image patches increases, posing a challenge for processing large images. Therefore, in underlying image processing applications that need to generate high-resolution outputs, recent approaches often employ different strategies to reduce the complexity. One potential remedy is to use the Swin transformer design to apply self-attention within localized image regions [28,29]. However, this design restricts contextual aggregation within the global neighbourhood, which goes against the primary motivation of using self-attention rather than convolution and,

thus, is not well suited for image restoration tasks. Inspired by Zamir et al.[30], we introduce an efficient transformer that models global information while maintaining computational efficiency.

III. APPROACH

The network structure of the proposed ET-MIMO-UNet is shown in Figure 1. The network mainly consists of a multiscale input encoder, an efficient transformer module(ETM), a multi-layer feature fusion block (MLFF), and a multiscale output decoder. Among them, two encoder blocks (EB) are used for the encoder, three decoder blocks (DB) are used for the decoder, and 12 efficient transformer layers (ETL) are used for the efficient transformer module (ETM). ET-MIMO-UNet is built on a single UNet architecture based on an encoder-decoder, which can fully utilize the multi-scale spatial features extracted from images by CNN. In addition, to capture global dependencies, an efficient transformer layer (ETL) is used to model global information, effectively combining local and global information to achieve multi-scale deblurring.

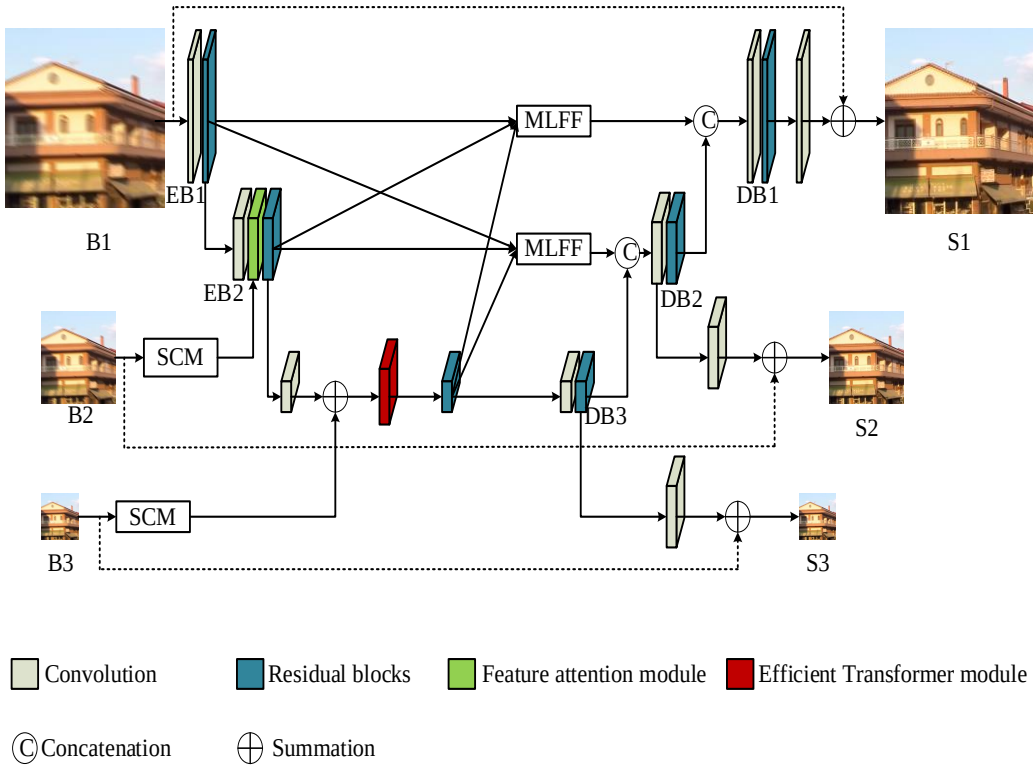

Figure 1: The Architecture of the Proposed Network

A. *Multi-scale Input Encoder*

Our proposed model incorporates a multi-scale input-output approach, which follows a coarse-to-fine strategy. This strategy has been widely adopted by various CNN-based deblurring models[10,16,31] and has proven its worth in terms of effectiveness. In the ET-MIMO-UNet encoder, different scales of blurry images are used as inputs through encoding blocks (EBs). EB1 consists of a convolutional layer and residual blocks; EB2 consists of a convolutional layer, a Feature Attention Module (FAM)[18] (See Figure 2), and residual blocks, and it has been proved that FAM can improve the performance of the model[18].

Using a multi-scale strategy as the input to a single U-Net, the original scale blurred image B1 is downsampled twice by 1/2, and the other two scales of blurred images B2 and B3 are obtained.EB1 and EB2 extract the blurred images of the B1 and B2 scales, and the blurred image of the B3 scale is fed into the Efficient Transformer Module (ELM) after preprocessing to perform global feature modelling. When extracting features from the encoder or transformer module at each scale, the Shallow Convolution Module (SCM) (See Figure 2)[18] is used to extract features from the downsampled images B2 and B3. SCM uses two stacked 3x3 and 1x1 convolutional layers. Then, in the last 1x1 convolutional layer, the extracted features are connected to the input current scale image, and another 1x1 convolutional layer is used to further refine the connection. The output of SCM is represented as Z_{out}^{k} .

For the original scale blurred image B1, do not use SCM; input it directly into the encoding block EB1. For the blurred image B2 using SCM, the Feature Attention Module (FAM) fuses the SCM output $Z_{out}^{2}$with the encoder output E_{out}^{1} at the B1 scale. Before fusion, a convolutional layer with a stride of 2 ensures that the two features

have the same size. Finally, a residual layer is used to refine the connection further. For blurred images at the B3 scale, the output Z_{out}^{3} of SCM is fused with the output $E_{out}^{2}$of the encoder at the B2 scale. After shallow feature extraction and convolutional feature extraction at the first two scales, each pixel has a deeper receptive field, which is fed into an efficient transformer module (ETM). This module utilizes the global information modeling ability of the transformer to establish long-term dependencies of features and then feeds the extracted global features into residual blocks.

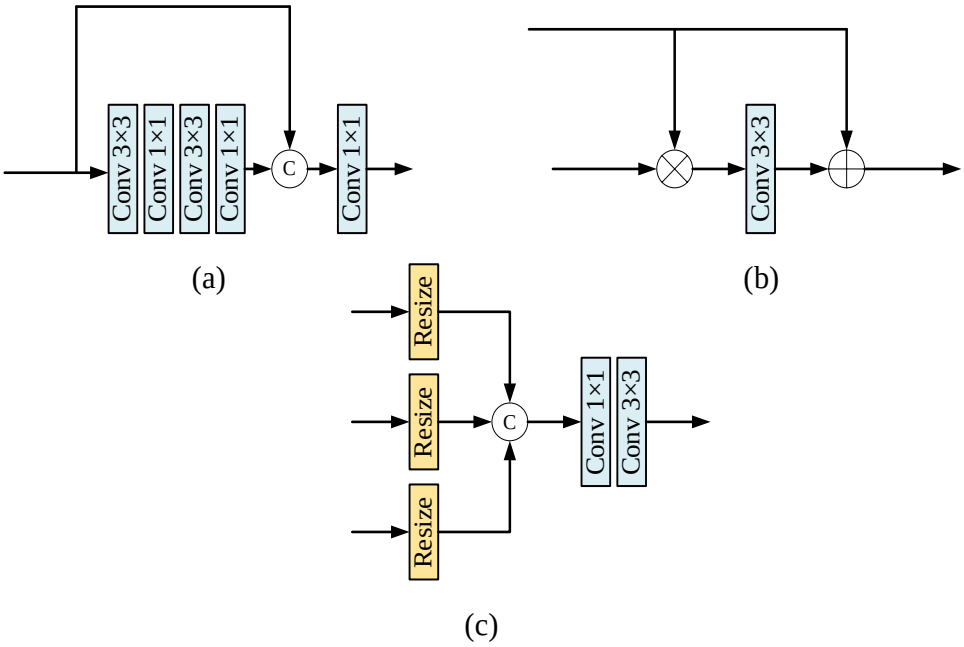

Figure 2: The Structures of Sub-modules: (a) SCM, (b) Feature Attention, and (c) MLFF.

B. Efficient Transformer Module

Although the Transformer is renowned for establishing long-term image dependencies, its high computational demands and substantial memory usage limit its widespread application. Consequently, handling high-resolution inputs becomes challenging. Inspired by Zamir et al.[30], we designed an efficient Transformer module (ETM) with the structure shown in Figure 3(a). Each ETM consists of multiple efficiency Transformer layers (ETL), whose structure is shown in Figure 3(b), and each ETL consists of multi-dconv head transposed attention (MHTA) and gated-dconv feed-forward network (GFFN). These two modules are described separately in the following.

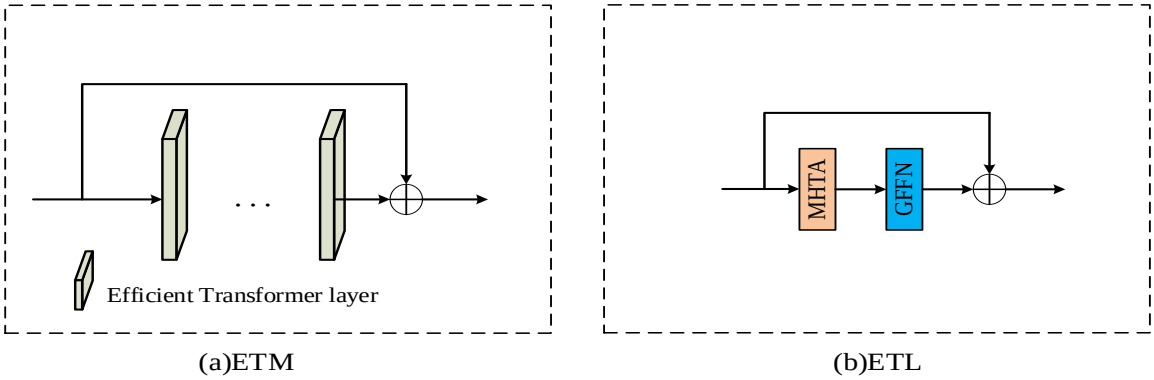

Figure 3: The Structures of ETM and ETL

1) Multi-dconv Head Transposed Attention: In Transformer, its computational complexity mainly comes from the self-attention layer, where the time and storage complexity of the dot product of query (Q) and key (K) in the traditional self-attention layer increase twice with the input spatial resolution. Therefore, applying self-attention layers to deblurring high-resolution images is not feasible. To solve this problem, this paper uses MHTA, which has linear complexity, as shown in Figure 4. The attention feature map of the self-attention layer is calculated across channels rather than spatial dimensions; that is, cross-channel covariance is calculated to generate an implicit encoding of the global context attention map. Before calculating feature covariance to generate attention feature maps, depth-wise convolution is introduced to emphasize local contextual information.

MHTA first generates a query (Q), key (K), and value (V) projection from the tensor Y that has been processed through layer normalization (LN). This process aggregates cross-channel pixels through 1x1 convolution, then applies 3x3 depth-wise convolution to encode channel-level contextual information. The formula is as follows:

$$Q = f_{3\times3}(f_{1\times1}(Y)) \quad (1)$$

$$K = f_{3\times3}(f_{1\times1}(Y)) \quad (2)$$

$$V = f_{3\times3}(f_{1\times1}(Y)) \quad (3)$$

Subsequently, it is necessary to capture global attention by establishing the correlation between Q and K. Firstly, reshape Q and K into matrix form, where $\tilde{Q} \in R^{C\times HW}$ and $\tilde{K} \in R^{HW\times C}$. Then, the dot product operation is applied for interaction to generate a transposed attention map of size $R^{C\times C}$. Then, V is also reshaped into matrix form, represented as $\tilde{V} \in R^{HW\times C}$, and the attention map is multiplied by $\tilde{V}$ to obtain the feature map $X_f \in R^{HW\times C}$. Finally, reshape X_f into tensor form, i.e. $\widetilde{X_f} \in R^{H\times W\times C}$, and apply 1x1 convolution operation to it. Before the softmax operation, we use temperature parameters β to control the dot product of Q and V. The appeal process is defined as follows:

$$X_f = Softmax(\tilde{Q} \cdot \tilde{K}/\beta) \cdot \tilde{V} \quad (4)$$

$$Y = f_{1\times1}(Reshape(X_f)) \quad (5)$$

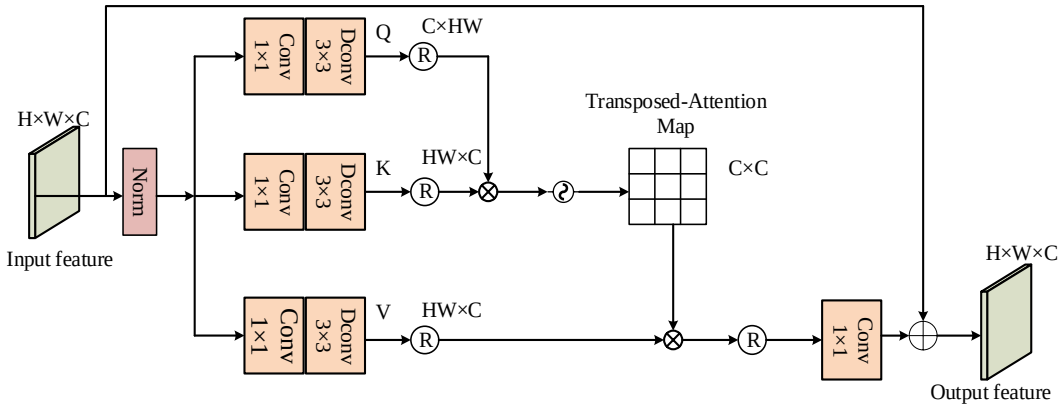

Figure 4: The Structures of MHTA

2) GFFN: GFFN is a network structure used to control information flow to suppress invalid information and preserve important information. In the GFFN shown in Figure 5, depth-wise convolution is also used. After layer normalization, adjacent pixel positions are encoded through depth-wise convolution to learn local image structures. Next, element-wise multiplication is used between the two branches to control the flow of information, with one branch activated by the GELU function. Finally, the output is obtained by refining the features through 1x1 convolution and adding input features. Assuming X represents the input feature, GFFN is defined as follows:

$$\tilde{X} = f_{3\times3}(f_{1\times1}(LN(X))) \quad (6)$$

$$Y = f_{1\times1}(\tilde{X} \cdot GELU(\tilde{X})) + X \quad (7)$$

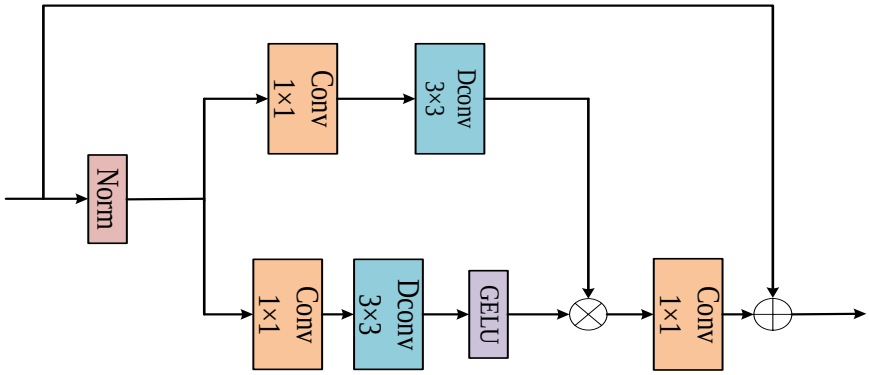

Figure 5: The Structure of GFFN

C. Multi-layer Feature Fusion Block

The multi-layer feature fusion block (MLFF) (See Figure 2) aims to fuse the features from the encoder output and the ETM output after the residual blocks, align their dimensions by up-sampling or down-sampling, and then splice them in channel dimensions and undergo convolutional fusion, and ultimately output them to the decoder to

facilitate the decoder to recover the image better and to reduce the loss of information. The detailed process can be formulated as follows:

$$MLFF_1^{out} = MLFF_1(EB_1^{out}, (EB_2^{out})^{\uparrow}, (ELM_{out})^{\uparrow}) \quad (8)$$

$$MLFF_2^{out} = MLFF_2((EB_1^{out})^{\downarrow}, EB_2^{out}, (ELM_{out})^{\uparrow}) \quad (9)$$

D. Multi-scale Output Decoder

After upsampling or downsampling, different scale features are fed into the MLFF module for feature fusion, and the fused features are fed to the network decoder for image reconstruction. The decoder still uses a single U-Net to simulate multiple cascaded U-Nets outputting deblurred images of different scales[18]. The decoding block (DB) consists of convolutional layers and residual blocks. Since the output of the decoding block is a feature map rather than an image, convolutional layers are used as mapping functions to generate deblurred images for each layer.

E. Loss Function

Adopting a coarse-to-fine approach, our model comprises three distinct stages, with each stage generating a progressively restored image. Therefore, we adopt a multi-scale loss to optimize our model, where the loss function combines two different losses: multi-scale content loss and multi-scale frequency reconstruction loss. Assume that G_K is the ground truth image of the K stage, and S_K is the corresponding restored image of the K stage, and t_k is the total number of elements of the K stage.

Multi-scale content loss: We use a multi-scale content loss function similar to other multi-scale networks[18]. The content loss function is defined as follows:

$$L_{cont} = \sum_1^K \frac{1}{t_k} \|S_K - G_K\|_1 \quad (10)$$

Studies have shown that adding auxiliary loss terms can improve model performance[18]. Auxiliary loss terms that minimize the distance between input and output in feature space have gained widespread adoption in image restoration tasks, leading to impressive outcomes[32]. For deblurring, it mainly restores the high-frequency components of the image[18]. As an auxiliary loss term, the multi-scale frequency reconstruction (MSFR) loss function is employed. This loss function calculates the L1 distance between the multi-scale ground-truth image and the restored image in the frequency domain. The specific definition is outlined below:

$$L_{MSFR} = \sum_1^K \frac{1}{t_k} \|FT(S_K) - FT(G_K)\|_1 \quad (11)$$

Here, F represents the fast Fourier transform (FFT), which converts the image signal into the frequency domain. The ultimate loss function for training our network is determined in the following manner, where λ is set 0.1:

$$L = L_{cont} + \lambda L_{MSFR} \quad (12)$$

IV. EXPERIMENTS

A. Datasets and Implementation Details

Train the network using the GoPro[10] training dataset, which includes 2103 pairs of blurred and clear images. The test dataset was comprised of GoPro, which includes 1111 pairs of images. In addition, to evaluate the generalization ability of the proposed model, the GoPro-trained network was directly applied to the ReaBlur[33] test dataset, which includes two sub-datasets: RealBlur-R and RealBlur- J, each containing 980 pairs of images.

We employed the Pytorch framework to train all models. During data preprocessing, we randomly cropped the images to a size of 256×256 and horizontally flipped them with a 50% chance. We conducted iterative training on the GoPro training dataset for 3000 epochs, with a batch size of 4. The initial learning rate for network training was set at 1×10-4, halved every 500 epochs. Additionally, our experiments were performed on a computer equipped with a GTX1080Ti GPU.

Taking into account the balance between computational efficiency and deblurring performance, we proposed variants of ET-MIMO-UNet, namely ET-MIMO-UNet+ and ET-MIMO-UNet++. Among them, 10 residual blocks for each EB and DB and 12 ETMs were used in ET-MIMO-UNet, and 20 residual blocks for each EB and DB and 6 ETMs were used in ET-MIMO-UNet+. ET-MIMO-UNet++ was based on ET-MIMO-UNet+ and replaced the depth-wise convolution in ETL with ordinary convolution.

B. Experimental Results

1) Quantitative Analysis: We tested all compared models under the same environment and evaluated image quality by peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics. Table 1 compares the

advanced models[9-11,13,16,18,19,31,34,35] based on the GoPro dataset. As can be seen from Table 1, compared with the MIMO-UNet basic network, the PSNR of the three models of ET-MIMO-UNet is improved by 0.41dB, 0.56dB, and 0.69dB, respectively; ET-MIMO-UNet is better than the comparison DeepDeblur, DeblurGAN, SRN, PSNR increased by 2.91dB, 3.44dB, 1.88dB respectively; Especially in the average SSIM, the method in this paper is significantly better than the comparison method.

Table 1: Deblurring Effect of the Advanced Deblurring Model onGoPro; The Optimal Model is Emphasized in Bold within the Table.

Method	PSNR(dB)	SSIM
FCL-GAN[19]	24.84	0.771
CRNet[13]	28.31	0.905
DeblurGAN[9]	28.70	0.858
DeepDeblur[10]	29.08	0.914
SRN[31]	30.26	0.934
PSS-NSC[16]	30.92	0.942
DMPHN[34]	31.20	0.945
Lian et el.[35]	31.53	0.948
CNBNet[11]	32.21	0.953
MIMO-UNet[18]	31.73	0.951
ET-MIMO-UNet(Ours)	32.14	0.958
ET-MIMO-UNet+(Ours)	32.29	0.959
ET-MIMO-UNet++(Ours)	32.42	0.960

We trained the model only on the GoPro dataset and tested it on the RealBlur dataset directly to evaluate the generalization ability of the proposed model. Table 2 shows the results of comparing the proposed method with the advanced method[9,10,18,19,31,34] on the RealBlur test dataset. As shown in Table 2, both PSNR and SSIM of our method outperform the comparison method, verifying the excellent generalization ability of ET-MIMO-UNet.

Table 2: Deblurring Results of Advanced Models on RealBlur are Evaluated Directly Using Gopro-trained Models; The Optimal Model is Emphasized in Bold within the Table.

Method	RealBlur-R		RealBlur-J	
	PSNR(dB)	SSIM	PSNR(dB)	SSIM
FCL-GAN[19]	28.37	0.663	25.35	0.736
DeepDeblur[10]	32.51	0.841	27.87	0.827
DeblurGAN[9]	33.79	0.903	27.97	0.834
SRN[31]	35.66	0.947	28.56	0.867
DMPHN[34]	35.70	0.948	28.42	0.860
MIMO-UNet[18]	35.47	0.946	27.76	0.836
ET-MIMO-UNet(Ours)	35.46	0.947	28.17	0.858
ET-MIMO-UNet+(Ours)	35.69	0.947	28.68	0.862
ET-MIMO-UNet++(Ours)	**35.74**	**0.948**	**28.72**	**0.868**

2) Qualitative Analysis: Figures 6 and 7 show examples of visual comparisons between our method and some existing algorithms on the GoPro test dataset and Real Blur test dataset, respectively. To present the restoration effect of each model more intuitively, we specially cropped out the local details of the image. Through comparison, it is clear that our proposed method performs better than other comparison methods in restoring image details.

Figure 6: Several Examples on the GoPro Test Dataset. From Top Left to Bottom Right: Blurry Images, Ground-truth Images, DeepDeblur[10], SRN[31], PSS-NSC[16], DMPHN[34], MIMO-UNet[18], ET-MIMI-UNet+(ours).

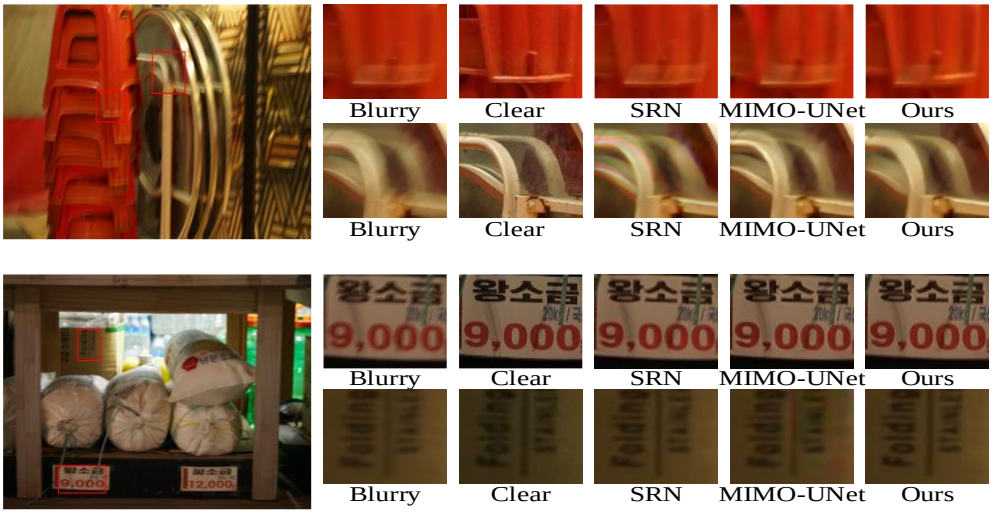

Figure 7: Several Examples on the RealBlur Test Dataset. From Left to Right: Blurry Images, Ground-truth Images, SRN[31], MIMO-UNet[18], ET-MIMI-UNet+(ours).

C. Ablation Study

We designed many experiments for the comprehensive evaluation of the proposed method.

1) To assess the effectiveness of each individual block, we conducted a series of experiments where we removed each block separately and trained the modified structures on the GoPro dataset. Our baseline model was a U-Net architecture with multiple inputs and outputs, excluding the ETM and MLFF components. Instead, it solely utilized 10 residual blocks in each encoder and decoder stage for feature extraction. We employed the same loss function described in Section 3.5 and adhered to the identical training approach. The outcomes of our experiments are summarized in Table 3.

Table 3: Impact of Various Components of ET-MIMO-UNet on the GoPro Test Dataset. The MLFF Indicates the Multi-layer Feature Fusion, and the ETM Indicates Efficient Transformer Module. The Symbol "√" Indicates that the Block is Included in the Training Process.

Basline	MLFF	ETM	PSNR(dB)	SSIM
√			31.55	0.946
√	√		31.74	0.952
√		√	31.96	0.953
√	√	√	32.14	0.958

As shown in Table 3, we comprehensively evaluated the PSNR and SSIM of ELM and MLFF. After introducing MLFF, PSNR increased by 0.19. It is worth noting that when ETM is introduced separately, the improvement of PSNR is more significant, reaching 0.41. The significant enhancement observed suggests that the global features captured by the Transformer module are pivotal in executing image deblurring tasks. By combining the global features of the Transformer with the local features of CNN, we significantly improved network performance.

2) To verify the superiority of the deblurring performance of the multi-scale strategy introduced in this paper, it was compared with the CNN-based single-scale deblurring model Deblur-GAN[12], SDWNet[32] and the dual-scale strategy-based deblurring model DeblurGAN-v2[33]. The results of comparing the GoPro test dataset are presented in Table 4.

Table 4: Comparison of Validity of Multi-scale Performance on the GoPro Test Dataset

Scale	Method	PSNR(dB)	SSIM
Sigle-scale	Deblur-GAN[18]	28.70	0.958
Sigle-scale	SDWNet[36]	31.26	0.966
Dual-scale	DeblurGAN-v2[37]	29.55	0.934
Multi-scale	ET-MIMO-UNet(ours)	32.14	0.958

As can be seen from Table 4, the multi-scale feature extraction method in this paper is better than the single-scale and dual-scale feature extraction methods, which verifies the advantages of multi-scale information extraction.

3) To delve deeper into the specific influence of the quantity of efficient Transformer layers (Num_ETL) within ETM on network performance, we undertook a comprehensive set of experiments utilizing the GoPro dataset. The comprehensive experimental findings are outlined in Table 5 below.

Table 5: Impact of the Number of ETLs on Model Performance. Num_ETL Means the Number of Efficient Transformer Layers (ETL).

NUM_ETL	PSNR(dB)	SSIM
0	31.74	0.952
4	31.88	0.953
8	31.96	0.955
12	32.14	0.958

When num_ETL=0, it means that the ETM proposed in this paper is not used. As shown in Table 5, after adding ELM, the model performance is significantly improved, and as the number of ETL increases, the network performance gradually increases. When the number of ETLs is greater than 12, the hardware used in this paper is difficult to train due to the increased complexity of the model. Therefore, this article finally chooses ETL=12 combined with CNNs as the final model. It is worth mentioning that experiments show that if hardware conditions permit, continuing to increase the number of ETLs in the model can achieve better PSNR and SSIM performance.

4) To prove the effectiveness of using depth-wise convolution in ETL to reduce model parameters, we replaced the depth-wise convolution in ETL with ordinary convolution and conducted experiments on the GoPro dataset. The experimental results are as follows:

Table 6: Validation of the Effectiveness of Depth-wise Convolution in ETL

Conv type	PSNR(dB)	SSIM	Params(M)
Conv	32.17	0.958	11.7
Depth-wise	32.14	0.958	11.1

As shown in Table 6, replacing depth-wise convolution with ordinary convolution network performance increases, but the parameters also increase accordingly. Considering the balance between network performance and network parameters, we used depth-wise convolution.

D. Execution Time and Parameters Comparison

In recent years, deblurring models have pursued a balance between accuracy, execution time, and parameters. Under similar precision, networks with smaller parameters are more suitable for deployment, and networks with less execution time are more popular. We focus on exploring an effective deblurring model with higher accuracy, lower parameters, and faster execution time. We tested all comparison models in the same environment. Table 7 compares our model's accuracy, model parameters, and processing speed with some state-of-the-art (SOTA) models[10,34,18,1].

Table 7: Parameter and Execution Time Comparison. Execution Time is the Average Execution Time Per Image on GoPro Test Data. All Models were Evaluated Using a Separate GTX1080Ti GPU.

	DeepDeblur[10]	DMPHN[34]	MIMO-UNet[18]	MPRNet[1]	ET-MIMO-UNet (ours)
Params (M)	11.7	21.7	6.8	20.1	11.1
Runtime (s)	4.330	0.424	0.013	1.976	0.257
PSNR (dB)	29.23	31.20	31.73	32.66	32.14
SSIM	0.916	0.945	0.951	0.959	0.958

As shown in Table 7, our model achieves PSNR and SSIM, which compete with other models and are relatively fast with relatively few parameters. It is worth noting that MPRNet has the highest PSNR and SSIM, but it has twice the number of parameters and eight times the processing speed of our model. MIMO-UNet has smaller PSNR and SSIM than our model, although it has fewer parameters and runs faster. Our model has a faster processing speed and better performance than DMPHN. This demonstrates that our model exhibits notable advantages in terms of model parameters, processing efficiency, and deblurring capabilities.

V. CONCLUSION

In this paper, we introduce a unique and innovative deblurring network. ET-MIMO-UNet, which effectively integrates Transformer into CNN-based UNet to realize single-image blind deblurring in dynamic scenes. This network not only inherits the strengths of CNN in modelling local contextual information but also leverages Transformer effectively to capture global semantic correlations. Experimental results on GoPro and RealBlur test datasets validate the effectiveness of the network model. In addition, our method exhibits better accuracy, faster speed, and smaller parameters than other methods. In the next step, we shall persist in exploring the integration of CNN and Transformer algorithms, aiming to refine the network model and enhance the deblurring capability in dynamic settings.

REFERENCES

[1] Zamir S W, Arora A, Khan S, et al. Multi-stage progressive image restoration//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14821-14831.

[2] Lou Y, Bertozzi A L, Soatto S. Direct sparse deblurring. Journal of Mathematical Imaging and Vision, 2011, 39: 1-12.

[3] Krishnan, Dilip, Terence Tay, and Rob Fergus. “Blind deconvolution using a normalized sparsity measure.” CVPR 2011. IEEE, 2011.

[4] Kotera, Jan, Filip Šroubek, and Peyman Milanfar. “Blind deconvolution using alternating maximum a posteriori estimation with heavy-tailed priors.” Computer Analysis of Images and Patterns: 15th International Conference, CAIP 2013, York, UK, August 27-29, 2013, Proceedings, Part II 15. Springer Berlin Heidelberg, 2013.

[5] Chakrabarti, Ayan. “A neural approach to blind motion deblurring.” Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016.

[6] Hradiš M, Kotera J, Zemcık P, et al. Convolutional neural networks for direct text deblurring//Proceedings of BMVC. 2015, 10(2).

[7] Sun J, Cao W, Xu Z, et al. Learning a convolutional neural network for non-uniform motion blur removal//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 769-777.

[8] Schuler C J, Hirsch M, Harmeling S, et al. Learning to deblur. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(7): 1439-1451.

[9] Kupyn O, Budzan V, Mykhailych M, et al. Deblurgan: Blind motion deblurring using conditional adversarial networks//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8183-8192.

[10] Nah S, Hyun Kim T, Mu Lee K. Deep multi-scale convolutional neural network for dynamic scene deblurring//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3883-3891.

[11] Zhang X, Zheng F, Jiang L, et al. CNB Net: A Two-Stage Approach for Effective Image Deblurring. Electronics, 2024, 13(2): 404.

[12] Lian, Zuozheng, Haizhen Wang, and Qianjun Zhang. “An Image Deblurring Method Using Improved U-Net Model.” Mobile Information Systems 2022 (2022).
[13] Zhao S, Zhang Z, Hong R, et al. Unsupervised color retention network and new quantization metric for blind motion deblurring. Authorea Preprints, 2023.
[14] Kong L, Dong J, Ge J, et al. Efficient frequency domain-based transformers for high-quality image deblurring//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5886-5895.
[15] Chen X, Wan Y, Wang D, et al. Image Deblurring Based on an Improved CNN-Transformer Combination Network. Applied Sciences, 2022, 13(1): 311.
[16] Gao H, Tao X, Shen X, et al. Dynamic scene deblurring with parameter selective sharing and nested skip connections//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3848-3856.
[17] Park D, Kang D U, Kim J, et al. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 327-343.
[18] Cho S J, Ji S W, Hong J P, et al. Rethinking coarse-to-fine approach in single image deblurring//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 4641-4650.
[19] Zhao S, Zhang Z, Hong R, et al. FCL-GAN: A lightweight and real-time baseline for unsupervised blind image deblurring//Proceedings of the 30th ACM International Conference on Multimedia. 2022: 6220-6229.
[20] Tsai F J, Peng Y T, Tsai C C, et al. BANet: A blur-aware attention network for dynamic scene deblurring. IEEE Transactions on Image Processing, 2022, 31: 6789-6799.
[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.
[22] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[23] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
[24] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers//European conference on computer vision. Cham: Springer International Publishing, 2020: 213-229
[25] .Ranftl, René, Alexey Bochkovskiy, and Vladlen Koltun. “Vision transformers for dense prediction.” Proceedings of the IEEE/CVF international conference on computer vision. 2021.
[26] Zeng, Yanhong, Jianlong Fu, and Hongyang Chao. “Learning joint spatial-temporal transformations for video inpainting.” Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16. Springer International Publishing, 2020.
[27] Yang F, Yang H, Fu J, et al. Learning texture transformer network for image super-resolution//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 5791-5800.
[28] Liang J, Cao J, Sun G, et al. Swinir: Image restoration using swin transformer//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1833-1844.
[29] Wang Z, Cun X, Bao J, et al. Uformer: A general u-shaped transformer for image restoration//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 17683-17693..
[30] Zamir S W, Arora A, Khan S, et al. Restormer: Efficient transformer for high-resolution image restoration//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5728-5739.
[31] Tao X, Gao H, Shen X, et al. Scale-recurrent network for deep image deblurring//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8174-8182.
[32] Zheng B, Yuan S, Slabaugh G, et al. Image demoireing with learnable bandpass filters//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3636-3645.
[33] Rim J, Lee H, Won J, et al. Real-world blur dataset for learning and benchmarking deblurring algorithms//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Springer International Publishing, 2020: 184-201.
[34] Zhang H, Dai Y, Li H, et al. Deep stacked hierarchical multi-patch network for image deblurring//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 5978-5986.
[35] Lian, Z., Wang, H. An image deblurring method using improved U-Net model based on multilayer fusion and attention mechanism. Sci Rep 13, 21402 (2023).
[36] Zou W, Jiang M, Zhang Y, et al. Sdwnet: A straight dilated network with wavelet transformation for image deblurring//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 1895-1904.
[37] Kupyn O, Martyniuk T, Wu J, et al. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 8878-8887.