Multi-scale Residual Low-Pass Filter Network for Image Deblurring

Jiangxin Dong Jinshan Pan* Zhongbao Yang Jinhui Tang
Nanjing University of Science and Technology

Abstract

We present a simple and effective Multi-scale Residual Low-Pass Filter Network (MRLPFNet) that jointly explores the image details and main structures for image deblurring. Our work is motivated by an observation that the difference between the blurry image and the clear one not only contains high-frequency contents[1] but also includes low-frequency information due to the influence of blur, while using the standard residual learning is less effective for modeling the main structure distorted by the blur. Considering that the low-frequency contents usually correspond to main global structures that are spatially variant, we first propose a learnable low-pass filter based on a self-attention mechanism to adaptively explore the global contexts for better modeling the low-frequency information. Then we embed it into a Residual Low-Pass Filter (RLPF) module, which involves an additional fully convolutional neural network with the standard residual learning to model the high-frequency information. We formulate the RLPF module into an end-to-end trainable network based on an encoder and decoder architecture and develop a wavelet-based feature fusion to fuse the multi-scale features. Experimental results show that our method performs favorably against state-of-the-art ones on commonly-used benchmarks.

1. Introduction

The camera shake and object motion are usually inevitable when taking photos with hand-held devices, which lead to significant motion blur effects. Restoring clear photos from blurry ones has attracted lots of attention from both research and industry communities [40]. However, motion deblurring is quite challenging as only the blurred images are available while the blur and latent images are unknown.

Significant progress has been made in image restoration due to the development of deep neural networks that directly learn the mapping from the degraded observation to the clear image. It is well known that the residual learning strategy [7] has been widely used to ease the training of networks, where a latent clear image I can be restored by the summation of the degraded input B and the output $\mathcal{N}(B)$ of a residual learning network $\mathcal{N}$. This strategy has been proven to be effective in lots of image restoration tasks, e.g., image super-resolution, as the degraded image B shares the same main structures as the clear one I, and the residual between B and I mainly contains image details.

However, we note that the difference between a blurry input and its corresponding clear one not only contains image details but also involves some structures. Taking Figure 1 as an example, the main structures in the clear image (e.g., the pillars enclosed in the red boxes of Figure 1(f)) disappear due to the influence of the motion blur effect. Thus, in addition to the details smoothed by the blur, some main structures also exist in the residual between the blurry image and the clear one (see the part enclosed in the red box of Figure 1(b)). As demonstrated in [22], the standard residual learning method [7] is effective for modeling the high-frequency information but less effective to restore the low-frequency contents. Figure 1(c) also shows that using the standard residual learning method may not estimate the main structures of the residual image well, which thus affects the final image restoration (Figure 1(h)). Therefore, it is important for image deblurring to effectively model the main structures distorted by the motion blur.

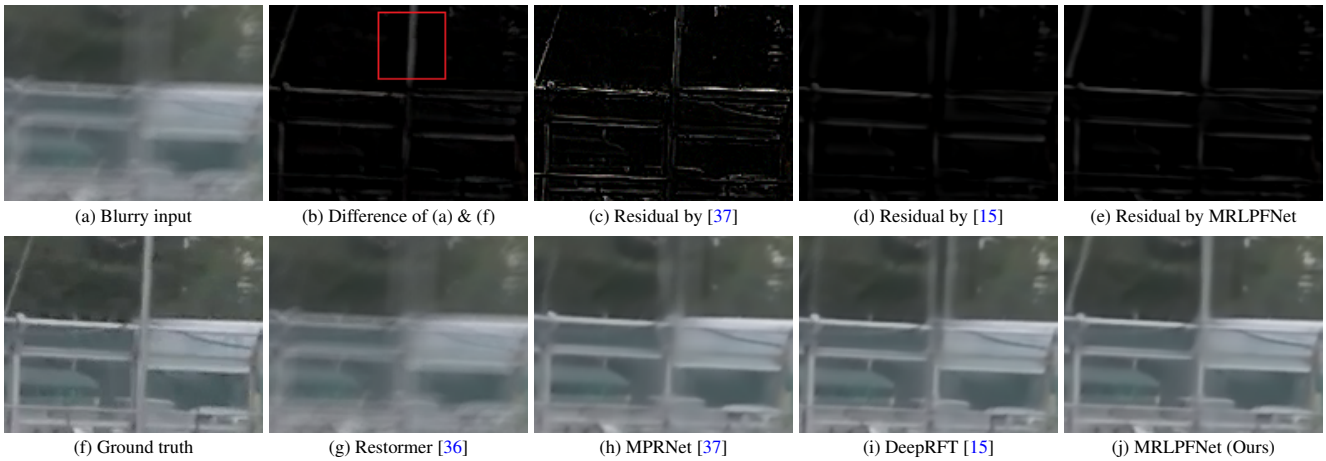
To alleviate the above-mentioned problem, the recent methods [15, 34] introduce a frequency branch based on the Fast Fourier Transform (FFT) in the residual network, which achieves good performance. However, as such a frequency branch is achieved by applying $\mathrm{Conv1} + \mathrm{ReLU} + \mathrm{Conv1}$[2] to the concatenation of the real and imaginary parts of the residual image, it may not effectively model the spatially-variant property of the global main structures in the residual image (Figure 1(d)), resulting in fake structures caused by the blur in the final deblurred image (Figure 1(i)).

In this paper, we propose a Multi-scale Residual Low-Pass Filter Network (MRLPFNet) for high-quality image deblurring. Our goal is to effectively model both the low-

*Corresponding author

[1]Note that the high-frequency contents in an image correspond to the image details, while the low-frequency ones denote the main structures of an image.

[2]$\mathrm{Conv1} + \mathrm{ReLU} + \mathrm{Conv1}$ denotes two 1×1 convolutional layers with the usual ReLU activation in between.



(a) Blurry input (b) Difference of (a) & (f) (c) Residual by [37] (d) Residual by [15] (e) Residual by MRLPFNet

(f) Ground truth (g) Restormer [36] (h) MPRNet [37] (i) DeepRFT [15] (j) MRLPFNet (Ours)

Figure 1. Visual comparison with state-of-the-art image deblurring methods on a challenging example. Our work is motivated by an observation that the difference between the blurry image and its corresponding clear one not only contains high-frequency contents but also includes low-frequency information due to the influence of the blur as shown in (b). We thus develop a simple yet effective MRLPFNet to learn both image details and spatially-variant structures (see (e)), which thus leads to better-deblurred results as shown in (j).

frequency and high-frequency parts of the residual image between the blurry input and its corresponding clear output within the residual learning framework. To better explore the low-frequency information, we propose to intuitively apply a low-pass filter on the residual image to concentrate on the low-frequency part. In addition, as the low-frequency contents in an image correspond to the global main structures, one of whose properties is that they are usually spatially variant, we exploit a learnable low-pass filter module based on a self-attention mechanism (as a basic module) to adaptively explore the global contexts so that the main structures can be better restored for image deblurring. Then we incorporate it in the proposed Residual Low-Pass Filter (RLPF) module with an additional residual learning branch based on a fully convolutional neural network (CNN) for modeling the high-frequency information.

Furthermore, the multi-scale strategy is widely used in image deblurring. As features from various scales have different spatial resolutions, resizing operations are usually used to fuse these features. However, simply using resizing operations based on downsampling or upsampling will lead to information loss. To this end, we develop a simple yet effective feature fusion module based on the wavelet transform to fuse the features with different spatial resolutions. Finally, we embed the proposed RLPF module and the wavelet-based feature fusion (WFF) module into an end-to-end trainable network within a coarse-to-fine framework. The proposed approach achieves favorable performance on widely-used benchmarks.

The main contributions are summarized as follows.

- We propose an effective RLPF module to model both the high-frequency and low-frequency information of the difference between the network input and output for image deblurring. Specifically, a learnable low-pass filter is developed based on a self-attention mechanism to adaptively explore the spatially-variant property of the global contexts to effectively reconstruct the low-frequency structures and a fully CNN with the residual learning is adopted to model the high-frequency information.
- We develop a simple yet effective feature fusion module based on the wavelet transform to fuse the features of different scales for better image deblurring.
- By training the proposed MRLPFNet in an end-to-end manner, we show that it performs favorably against state-of-the-art methods.

2. Related Work

Hand-crafted prior-based methods. As image deblurring is ill-posed, conventional methods usually develop kinds of effective priors to constrain the solution space, e.g., sparsity priors on the image gradient [11, 35] or intensity [19], internal patch recurrence [16], sparsity of dark channel prior [21], etc. These hand-crafted priors can help the blur removal. However, they do not fully exploit the characteristics of the clear image data and usually lead to complicated optimization problems.

Deep learning-based methods. Instead of manually designing image priors, lots of methods develop kinds of deep CNNs to solve image deblurring. Early approaches employ deep CNN models to estimate blur kernels and then use existing prior-based restoration methods to estimate the clear images [27, 6, 1, 24]. While inaccurate blur kernels will lead to deblurred results with severe artifacts as the blur kernels and latent clear images are estimated independently.

Instead of estimating blur kernels, numerous methods [17, 39, 28, 5, 38, 26, 37, 3, 18, 12] directly estimate

clear images from blurry ones. In [17], Nah et al. develop a multi-scale deep CNN, where the restored images from coarse scales are resized to help the estimation of finer scales. To better explore the multi-scale information, Tao et al. [28] develop an effective scale recurrent neural network. In [5], Gao et al. propose a parameter selective sharing and nested skip connection method to improve [17]. Although the methods based on a multi-scale strategy achieve decent performance, simply increasing the model depth with more scales cannot further improve the quality of deblurring. To overcome this problem, methods based on multi-patches have been developed. Zhang et al. [38] develop an effective deep hierarchical multi-patch network, where the features from the previous stage are concatenated to facilitate the estimation of the following stages. Zamir et al. [37] propose a multi-stage progressive method to better explore the features from multi-stages by a cross-scale feature fusion module. In addition, an effective supervised attention module is developed to better guide the estimations of the following stages. As both the multi-scale and multi-patch methods need to perform the network at each scale or stage recurrently, which leads to high computational costs, Cho et al. [3] develop a multi-input and multi-output U-net based on a multi-scale strategy to solve image deblurring.

These above-mentioned methods achieve significant performance, most of which adopt the residual learning in the network design. As demonstrated in [20], the residual learning is able to model the image details (i.e, high-frequency information), but is less effective for exploring the main structures (i.e, low-frequency information). To overcome this problem, Mao et al. [15] improve the residual learning network by introducing an additional branch performed in the frequency domain. However, as the main structures involve the spatially-variant global contexts, simply using the local operations (e.g., 1×1 convolution in [15]) may not model the main structures well.

Transformer-based methods. As Transformer is effective for global context exploration and shows great potential in many vision tasks, several methods apply it to image deblurring. Zamir et al. [36] propose an efficient Transformer model by estimating self-attention along the channel dimension. This method is further simplified by [2]. Tsai et al. [29] develop a Transformer-based model by constructing intra- and inter-strip tokens to reweight image features in the horizontal and vertical directions. In [30], Tu et al. develop an effective multi-axis MLP-based architecture. Wang et al. [31] develop a general U-shaped Transformer to solve image deblurring. These Transformer-based approaches achieve decent results. However, as our analysis in Section 3 shows, self-attention has the same effect as the low-pass filter. Simply using Transformers may lead to over-smoothed results.

3. Low-Pass Filter and Self-Attention

To better motivate our work, we first describe the relations between the low-pass filter and self-attention.

Given an image I and a low-pass filter $\mathcal{F}$, we can obtain a filtered image f by:

$$f(\mathrm{x}) = \mathcal{F}(I(\mathrm{x})) = \sum_{\mathrm{y}} \mathbf{W}_{\mathrm{xy}} I(\mathrm{y}), \tag{1}$$

where x and y denote image pixels; $\mathbf{W}_{\mathrm{xy}}$ denotes the filter weight for $\mathcal{F}$, satisfying $\mathbf{W}_{\mathrm{xy}} \geq 0$ and $\sum_{\mathrm{y}} \mathbf{W}_{\mathrm{xy}} = 1$. The role of a low-pass filter is to do a weighted average in an area with the same size as the filter, thus filtering out the high-frequency information and retaining the low-frequency one. One common type of filter is the linear translation-invariant filter, e.g., the Gaussian filter, which adopts the spatially-invariant filter weight and cannot effectively describe the variant properties of image structures in different spatial areas. Another type of filter is the non-linear filter, e.g., the bilateral filter, whose weights depend on both the spatial closeness and the intensity difference. Such weights can adaptively model the spatially-variant contexts (e.g., global edges), thus preserving more useful low-frequency information in the filtered image.

Recently, the attention-based approaches, e.g., self-attention, compute the weighted value as

$$\mathbf{Z} = \mathbf{S} \circledast \mathbf{V} = \mathrm{softmax}\left(\frac{\mathcal{Q}(F)\mathcal{K}(F)^{\top}}{\sqrt{d_k}}\right)\mathcal{V}(F), \tag{2}$$

where F denotes the deep feature extracted from the input image I; $\circledast$ indicates the matrix multiplication; $\mathcal{Q}(\cdot)$, $\mathcal{K}(\cdot)$, and $\mathcal{V}(\cdot)$ denote the operations that extract the matrices for the query, key, and value; d_k is the dimension of the keys. A softmax normalized operation is applied to each row of $\mathbf{S}$ to ensure $\mathbf{S}_{ij} \geq 0$ and $\sum_j \mathbf{S}_{ij} = 1$, where i and j denote the index of the row and column in $\mathbf{S}$. Thus, each row of $\mathbf{S}$ can be regarded as a low-pass filter. Compared to the bilateral filter whose weight measures the spatial and color similarity based on the exponential Euclidean distance, the attention value measures the similarity of query and key based on the correlation metric. Therefore, it is spatially variant and can model the main structures more adaptively.

Considering the self-attention (2) is learnable and more flexible, we propose to straightforwardly embed it into the network to better model the low-frequency information for image deblurring.

4. MRLPFNet

Our goal is to develop an effective MRLPFNet for high-quality image deblurring, which can better restore both the image details and main structures. Specifically, we first propose a learnable low-pass filter based on the self-attention method to model the low-frequency information. We then

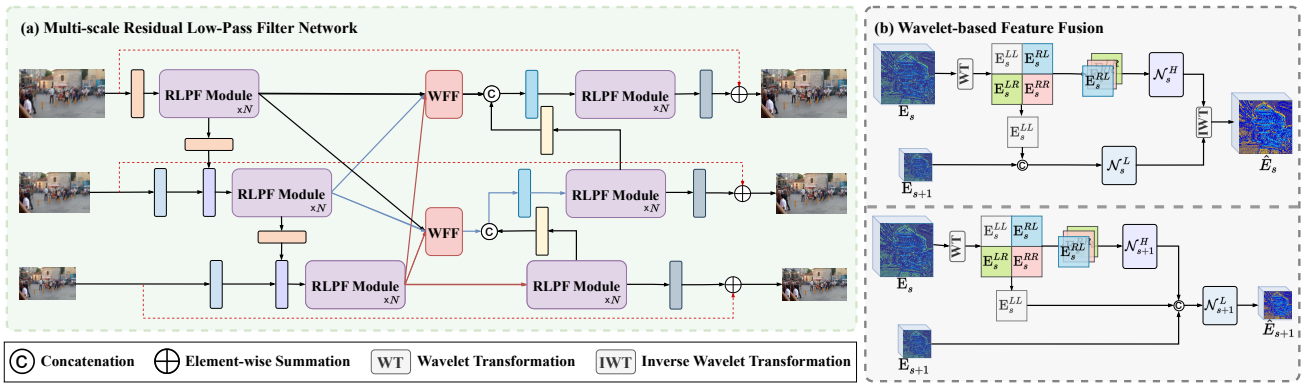


Figure 2. Multi-scale residual low-pass filter network. To restore high-quality images with clearer structures, we propose an effective RLPF module that involves a learnable low-pass filter (based on the self-attention) to explore the low-frequency information (through Eqs. (3), (4), and (5)). A wavelet-based feature fusion (WFF) module is developed to fuse the features of various scales for better image deblurring.

incorporate it in our proposed RLPF module with an additional residual learning branch for high-frequency information exploration. Furthermore, we formulate it into an end-to-end trainable network based on a coarse-to-fine framework and propose a feature fusion module based on the wavelet transform to better utilize the multi-scale features. In the following, we present the details of the proposed method.

4.1. Residual Low-Pass Filter

As the standard residual learning is adept at exploring the high-frequency information [9, 20] while less effective for the low-frequency information modeling, the method [15] improves the standard residual block by introducing an additional branch that applies $\mathrm{Conv1} + \mathrm{ReLU} + \mathrm{Conv1}$ to the features (generated by the FFT) in the frequency domain. Although decent performance has been achieved, the 1×1 convolution is spatially invariant, which does not effectively model the spatially-variant structures. To overcome this problem, we propose to further improve it by applying a learnable low-pass filter to concentrate on the global main structures, which is able to model the spatially-variant property of the low-frequency information.

Specifically, we first generate the feature $\mathbf{Y}$ from a blurry input I by a shallow feature extraction network (the details are included in the supplemental material). Then the proposed RLPF module explores the properties of residual learning in both the spatial and frequency domains according to [15]:

$$\mathbf{F} = \mathbf{Y} + \mathcal{R}(\mathbf{Y}) + \mathcal{F}(\mathbf{Y}), \tag{3}$$

where $\mathcal{R}(\cdot)$ denotes a plain network that applies $\mathrm{Conv3} + \mathrm{ReLU} + \mathrm{Conv3}$[3] in the spatial domain and $\mathcal{F}(\cdot)$ denotes a network [15] with $\mathrm{Conv1} + \mathrm{ReLU} + \mathrm{Conv1}$ employed in the frequency domain.

[3] $\mathrm{Conv3} + \mathrm{ReLU} + \mathrm{Conv3}$ denotes two 3×3 convolutional layers with the usual ReLU activation in between.

As Eq. (3) is based on the locally spatially-invariant convolution operations, it does not effectively model the spatially-variant main structures (see Figure 1(d)). To this end, we develop a learnable low-pass filter, which is proposed based on Eq. (1) and achieved by:

$$\mathbf{h}_i = \sum_j \mathbf{S}_{ij}\mathbf{f}_j, \tag{4}$$

where $\mathbf{f}_j$ is the j-th feature map of $\mathbf{F}$. Based on the discussions in Section 3, the self-attention can be regarded as a spatially-variant low-pass filter, which can better model the global low-frequency contents. Then we use the standard scaled dot-production attention in the Transformer according to [36] to estimate the filter weight matrix $\mathbf{S}$. Finally, we obtain the filtered feature $\mathbf{E}$ by:

$$\mathbf{E} = \mathbf{F} + \mathcal{D}(\mathbf{H}), \tag{5}$$

where $\mathbf{H} = \{\mathbf{h}_i\}$ is the low-frequency feature by Eq. (4) and $\mathcal{D}$ denotes a 3×3 depth-wise convolution layer.

We refer to the module achieved by Eqs. (3), (4), and (5) as the proposed RLPF module, which is then embedded into an encoder-decoder architecture as shown in Figure 2(a).

4.2. Wavelet-based Feature Fusion

To efficiently ease and improve the deblurring process, the recent methods [3, 15] utilize the multi-scale features generated by different encoder blocks to prompt the learning of various decoder blocks. As the spatial resolutions of the features from different scales are not the same, they are usually resized by downsampling or upsampling operations to the same spatial resolution for feature fusion. However, these resizing operations may lose some important structural details, thus affecting the final image restoration. As the wavelet transform [8] can model the image scale information and is insertable, we develop a wavelet-based feature fusion method to better fuse the features of various scales for image deblurring.

Specifically, considering two features $\mathbf{E}_s$ and $\mathbf{E}_{s+1}$ generated by Eq. (5) for the finer image scale s and coarser image scale $s+1$, we first apply the Haar transform $\mathcal{H}$ [32, 14, 33] to $\mathbf{E}_s$ and obtain

$$\{\mathbf{E}_s^{LL}; \mathbf{E}_s^{LR}; \mathbf{E}_s^{RL}; \mathbf{E}_s^{RR}\} = \mathcal{H}(\mathbf{E}_s), \tag{6}$$

where $\mathbf{E}_s^{LL}$ denotes the low-frequency part and $\mathbf{E}_s^{LR}$, $\mathbf{E}_s^{RL}$, and $\mathbf{E}_s^{RR}$ correspond to three high-frequency parts. Thus, the resolutions of $\{\mathbf{E}_s^{LL}; \mathbf{E}_s^{LR}; \mathbf{E}_s^{RL}; \mathbf{E}_s^{RR}\}$ are the same as that of $\mathbf{E}_{s+1}$. To make use of both $\mathbf{E}_s$ and $\mathbf{E}_{s+1}$ for the decoder block in the finer image scale s, the wavelet-based feature fusion generates the fused feature $\hat{\mathbf{E}}_s$ by:

$$\hat{\mathbf{E}}_s = \mathcal{H}^{-1}\left(\mathcal{N}_s^L(\mathcal{C}(\mathbf{E}_s^{LL}, \mathbf{E}_{s+1})), \mathcal{N}_s^H(\mathcal{C}(\mathbf{E}_s^{LR}, \mathbf{E}_s^{RL}, \mathbf{E}_s^{RR}))\right), \tag{7}$$

where $\mathcal{H}^{-1}$ is the inverse Haar transform; $\mathcal{C}(\cdot)$ indicates the concatenation along the channel dimension; $\mathcal{N}_s^L$ is a network consisting of two 3×3 convolutional layers with the PReLU activation in between and $\mathcal{N}_s^H$ contains one 1×1 convolutional layer followed by the PReLU activation and one 3×3 convolutional layer. When fusing $\mathbf{E}_s$ and $\mathbf{E}_{s+1}$ to facilitate the decoder block in the coarser image scale $s+1$, we obtain the fused feature $\hat{\mathbf{E}}_{s+1}$ by

$$\hat{\mathbf{E}}_{s+1} = \mathcal{N}_{s+1}^L\left(\mathcal{C}(\mathbf{E}_{s+1}, \mathbf{E}_s^{LL}, \mathcal{N}_{s+1}^H(\mathcal{C}(\mathbf{E}_s^{LR}, \mathbf{E}_s^{RL}, \mathbf{E}_s^{RR})))\right), \tag{8}$$

where $\mathcal{N}_{s+1}^L$ and $\mathcal{N}_{s+1}^H$ have the same network architecture as $\mathcal{N}_s^L$ and $\mathcal{N}_s^H$, respectively (Note that the network parameters are not shared across scales). We take the fused features $\hat{\mathbf{E}}_s$ and $\hat{\mathbf{E}}_{s+1}$ instead of $\mathbf{E}_s$ and $\mathbf{E}_{s+1}$ as the input for the following decoder blocks. Figure 2(b) shows the details of the proposed wavelet-based feature fusion module.

When fusing features from more scales, we can recurrently use Eqs. (7) and (8). As the wavelet transform is invertible, all the information can be better preserved during the transformation. We show the effectiveness of the wavelet-based feature fusion in Section 6.2.

5. Experimental Results

We first describe the implementation details of the proposed MRLPFNet. Then we evaluate our approach on the commonly used benchmarks and compare it against the state-of-the-art methods. Due to the page limit, we include more results in the supplemental material. The PyTorch code and trained models are available at our Project page.

5.1. Implementation details

We embed the proposed RLPF module into an encoder-decoder network [3], which contains three encoder blocks and three decoder blocks. For each encoder/decoder block, we adopt a stack of 8 RLPF modules (i.e., $N = 8$ in Figure 2(a)). We then use the wavelet-based feature fusion module to fuse the features generated by the encoder blocks of 3 different image scales. We implement our method based on the PyTorch framework and train it from scratch using a machine with two NVIDIA GeForce RTX 3090 GPUs. The proposed network is trained using the Adam optimizer [10] with default parameters. The batch size is set to be 8. The size of each image patch is 256×256 pixels. The learning rate is initialized to be 2×10^{-4} and is updated by the Cosine Annealing scheme. The loss function [37] is adopted to constrain the network training.

Table 1. Quantitative evaluations of the proposed approach against state-of-the-art methods on the GoPro dataset [17]. "MRLPFNet-L" denotes the proposed method using 20 RLPF modules in each encoder and decoder block (i.e., $N = 20$ in Figure 2(a)).

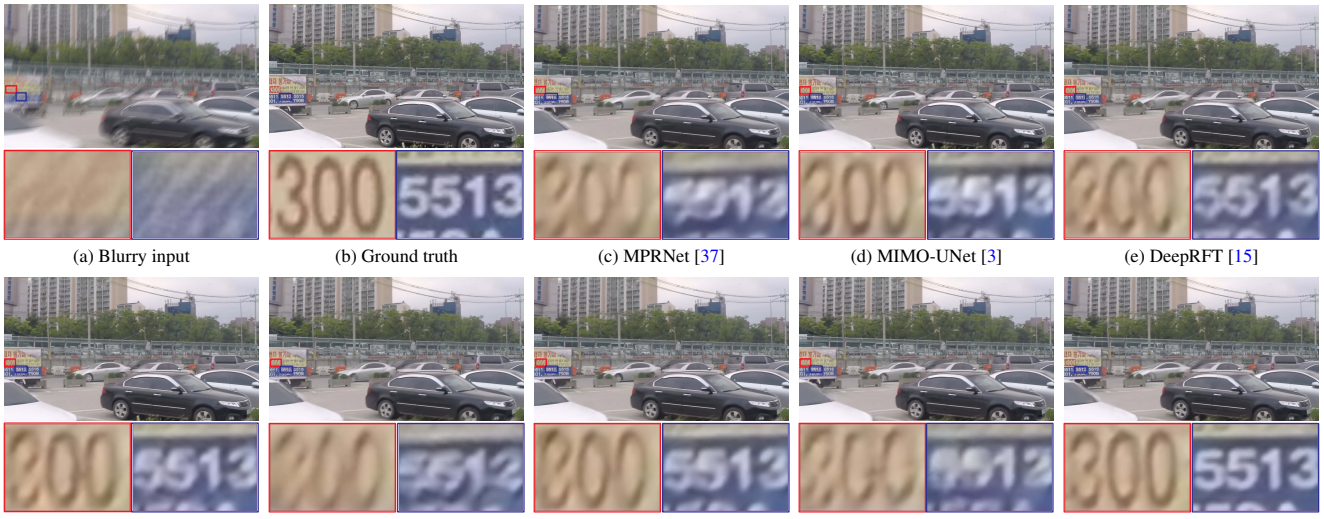
Method	PSNR (dB)	SSIM
SRN [28]	30.26	0.9342
SSN [5]	30.92	0.9421
DMPHN [38]	31.20	0.9453
SAPHNet [26]	31.85	0.9480
MPRNet [37]	32.66	0.9589
MPRNet-Local [4]	33.31	0.9637
MIMO-UNet [3]	31.73	0.9510
DeepRFT [15]	32.82	0.9600
Uformer [31]	33.06	0.9670
Restormer [36]	32.92	0.9611
Restormer-Local [4]	33.57	0.9656
MAXIM [30]	32.86	0.9616
NAFNet [2]	33.71	0.9668
Stripformer [29]	33.08	0.9624
MRLPFNet	33.50	0.9650
MRLPFNet-L	**34.01**	**0.9682**

5.2. Comparisons with the state of the art

Evaluations on the GoPro dataset. We first quantitatively evaluate the proposed method on the GoPro dataset [17], which contains 2,103 images for training and 1,111 images for the test. We compare the proposed method against several state-of-the-art methods including CNN-based methods (SRN [28], SSN [5], DMPHN [38], SAPHNet [26], MPRNet [37], MIMO-UNet [3], DeepRFT [15], NAFNet [2]), Transformer-based methods (Restormer [36], Uformer [31], Stripformer [29]), and MLP-based methods (MAXIM [30]). For fair comparisons, we fine-tune or retrain the deep learning-based methods that are not trained on the GoPro dataset. We use the PSNR and SSIM as metrics to evaluate the quality of restored images.

Table 1 shows the quantitative results. The proposed method generates higher-quality images with higher PSNR and SSIM values than the competing approaches.

Figure 3 shows a visual comparison. Most of the CNN-based methods, e.g, [37, 3, 2], employ the commonly used residual learning in the network designs, which is less effective for the estimation of the main structures, as demonstrated in Section 1. Thus, these methods [37, 3, 2] do not effectively remove the blur as shown in Figure 3(c), (d), and



(a) Blurry input (b) Ground truth (c) MPRNet [37] (d) MIMO-UNet [3] (e) DeepRFT [15]

(f) NAFNet [2] (g) Restormer [36] (h) Restormer-Local [4] (i) Stripformer [29] (j) MRLPFNet

Figure 3. Example from the GoPro dataset [17]. Compared to the evaluated methods in (c)-(i), the proposed approach generates a better-deblurred image with clearer characters in (j).

(f). The method [15] improves the residual learning by introducing an additional branch operating in the frequency domain. However, as the main structures are usually spatially variant, simply using localized convolution operations does not effectively model such global properties. The deblurred result in Figure 3(e) still contains the blur effect.

The recent methods [36, 29] develop Transformers to solve image deblurring. As analyzed in Section 3, Transformers can be regarded as low-pass filters. These methods are able to explore the low-frequency contents but less effective for modeling the high-frequency information, thus leading to over-smoothed deblurred results as shown in Figure 3(g) and (i). In contrast to the above-mentioned methods, the proposed MRLPFNet is able to explore both high- and low-frequency information, generating a better image with clearer details and structures than the competed methods (e.g., the restored numbers in Figure 3(j) are clearer).

Evaluations on the RealBlur dataset. The RealBlur dataset [23] contains 3,758 image pairs for training and 980 image pairs for the test. Based on the protocols of [23], our approach is compared with the state-of-the-art methods for fair comparisons. Table 2 shows that the proposed method generates the deblurred results with higher PSNR and SSIM values, where the PSNR values of our method are at least 1.08dB and 0.71dB higher than the evaluated methods on the "RealBlur-R" and "RealBlur-J" datasets, respectively.

Figure 4 shows some visual comparisons of the evaluated methods. The methods [28, 13, 3] based on the standard residual learning do not effectively remove the blur effects as shown in Figure 4(c)-(e). The method [15] improves the standard residual learning, but the restored images are over-smoothed, e.g., the strokes in Figure 4(f). The Transformer-based method [29] tends to generate over-smoothed results and does not effectively remove the blur as shown in Figure 4(g). In contrast, our method generates clearer images than the evaluated methods (Figure 4(h)).

Table 2. Quantitative evaluations of the proposed approach against state-of-the-art methods on the RealBlur dataset [23]. All the comparison results are generated using the publicly available codes and the models trained on the same training datasets.

	RealBlur-R		RealBlur-J	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
SRN [28]	38.65	0.9652	31.38	0.9091
DeblurGAN [12]	36.44	0.9347	29.69	0.8703
MIMO-UNet+ [3]	-	-	31.92	0.9190
DeepRFT+ [15]	39.84	0.9721	32.19	0.9305
Stripformer [29]	39.84	0.9737	32.48	0.9290
MRLPFNet	**40.92**	**0.9753**	**33.19**	**0.9361**

Table 3. Quantitative evaluations of the proposed approach against state-of-the-art methods on the HIDE dataset [25]. All the comparison results are generated by the same models trained on the GoPro dataset [17] as in Table 1.

Method	PSNR (dB)	SSIM
SRN [28]	28.36	0.9040
SAPHNet [26]	29.98	0.9300
MPRNet [37]	30.96	0.9394
MPRNet-Local [4]	31.19	0.9418
MIMO-UNet [3]	29.28	0.9206
MIMO-UNet+ [3]	29.99	0.9304
DeepRFT [15]	30.99	0.9407
DeepRFT+ [15]	31.42	0.9442
Restormer [36]	31.22	0.9423
Restormer-Local [4]	31.49	0.9447
Stripformer [29]	31.03	0.9395
MRLPFNet	**31.63**	**0.9465**

(a) Blurry input (b) Ground truth (c) SRN [28] (d) DeblurGAN-V2 [13]

(e) MIMO-UNet [3] (f) DeepRFT+ [15] (g) Stripformer [29] (h) MRLPFNet

Figure 4. Example from the RealBlur dataset [23]. Compared with the deblurred results in (c)-(g), the proposed method recovers a high-quality image with clearer structural details.

Table 4. Effectiveness of the proposed low-pass filter and wavelet-based feature fusion for image deblurring. All baseline methods are trained using the same settings as the proposed method for fair comparisons.

	Low-pass filter (LPF)	Feature fusion (FF)		GoPro
		Bilinear-based (BFF)	Wavelet-based (WFF)	PSNR (dB)/SSIM
$\text{MRLPFNet}_{\text{w/o LPF\&FF}}$	✗	✗	✗	32.8/0.9606
$\text{MRLPFNet}_{\text{w/o LPF\&w/ BFF}}$	✗	✓	✗	32.9/0.9611
$\text{MRLPFNet}_{\text{w/o LPF\&w/ WFF}}$	✗	✗	✓	33.0/0.9616
$\text{MRLPFNet}_{\text{w/ LPF\&w/o FF}}$	✓	✗	✗	33.4/0.9641
MRLPFNet	✓	✗	✓	**33.5/0.9650**

Evaluations on the HIDE dataset. We further evaluate our method on the HIDE dataset [25] by using the model trained on the GoPro dataset. Table 3 shows that the proposed approach generates the deblurred results with the highest PSNR and SSIM values.

6. Analysis and Discussions

6.1. Effect of the learnable low-pass filter

The learnable low-pass filter is proposed to explore the spatially-variant low-frequency information so that the global structural contents can be better restored for image deblurring. To demonstrate the effectiveness of the proposed learnable low-pass filter, we first compare our method with a baseline that removes the learnable low-pass filter ($\text{MRLPFNet}_{\text{w/o LPF\&w/ WFF}}$ for short) and train this baseline using the same settings as ours. We use the GoPro dataset [17] as described in Section 5.2 for evaluation. Table 4 shows the quantitative results, where the PSNR of the proposed MRLPFNet is 0.5dB higher than the baseline without the learnable low-pass filter. To focus on the effect of the low-pass filter, we further disable the feature fusion module and compare the baselines with and without the learnable low-pass filter ($\text{MRLPFNet}_{\text{w/o LPF\&FF}}$ and $\text{MRLPFNet}_{\text{w/ LPF\&w/o FF}}$ for short). The comparisons in Table 4 illustrate the significance of using the learnable low-pass filter, which is able to concentrate on the low-frequency information modeling and effectively recover the spatially-variant image structures for image deblurring.

Figure 5 shows a visual comparison. The main structures

(a) Blurry input (b) MRLPFNet$_{\text{w/o LPF\&w/ WFF}}$ (c) MRLPFNet

(d) (e) (f)

Figure 5. Effect of the proposed learnable low-pass filter. (d) is the residual image between the blurry image and Ground truth. (e) and (f) are the residual images estimated by MRLPFNet$_{\text{w/o LPF\&w/ WFF}}$ and MRLPFNet, respectively.

in the residual image estimated by MRLPFNet$_{\text{w/o LPF\&w/ WFF}}$ are not well recovered, e.g., the license plate numbers in Figure 5(e), resulting in the deblurred image with a blur effect (Figure 5(b)). In contrast, the residual image estimated by the proposed MRLPFNet is better restored, as using the learnable low-pass filter is able to model the spatially-variant image structures. Thus, our approach is more effective at yielding a clear image with distinct structures (e.g., the license plate numbers in Figure 5(f)).

6.2. Effect of the wavelet-based feature fusion

To demonstrate the effectiveness of the wavelet-based feature fusion on fusing the features with different spatial resolutions, we first disable the learnable low-pass filter in our implementation (MRLPFNet$_{\text{w/o LPF\&w/ WFF}}$ for short). Then we compare with the baseline that replaces the wavelet-based feature fusion with the bilinear interpolation (MRLPFNet$_{\text{w/o LPF\&w/ BFF}}$ for short). The results in Table 4 show that using the bilinear interpolation as the feature fusion does not perform well as our method with the wavelet-based one, since the wavelet transform is invertible and can preserve more accurate information during the transform process. The qualitative comparisons in Figure 6 further show that using the wavelet-based feature fusion is able to improve the quality of the restored image (Figure 6(c)). We also compare our complete network with the baseline that removes the feature fusion module (MRLPFNet$_{\text{w/ LPF\&w/o FF}}$ for short). The comparisons in Figure 6 show the same tendency that using the wavelet-based feature fusion improves the final deblurring performance.

6.3. Model complexity

We examine the model complexity of the proposed approach and the state-of-the-art methods in terms of model parameters, running time, and FLOPs. Table 5 shows that our method has the lowest FLOPs value and also achieves competitive performance against the competed ones in terms of running time.

(a) Blurry input (b) (c) (d) Ground truth

Figure 6. Effect of the proposed wavelet-based feature fusion. (b) and (c) are the results obtained by the baselines using the Bilinear-based (MRLPFNet$_{\text{w/o LPF\&w/ BFF}}$ in Table 4) and the wavelet-based (MRLPFNet$_{\text{w/o LPF\&w/ WFF}}$ in Table 4) feature fusion, respectively.

Table 5. Model complexity of the top-performance methods on the above-mentioned datasets. The FLOPs are evaluated on image patches with the size of 256×256 pixels based on the protocols of existing methods. The running time is evaluated on images with the size of 1280×720 pixels. All the results are obtained on a machine with an NVIDIA GeForce RTX 3090 GPU.

Methods	Model parameters (M)	FLOPs (G)	Running time (/s)
MPRNet [37]	20.1	760	0.9806
MIMO-UNet+ [3]	16.1	151	0.2827
DeepRFT+ [15]	23.0	183	0.5518
Restormer [36]	26.1	155	0.9798
Stripformer [29]	19.7	170	0.6580
MRLPFNet$_{\text{w/o LPF}}$	11.4	82	0.2613
MRLPFNet	20.6	129	0.8633

Limitations. As the proposed learnable low-pass filter is based on the self-attention to estimate the filter weights, it moderately increases the running time as shown in Table 5 (see the comparisons of "MRLPFNet$_{\text{w/o LPF}}$" and "MRLPFNet"). Future work will develop efficient and learnable low-pass filters for better image deblurring.

7. Conclusion

We present an effective MRLPFNet to reconstruct both the low-frequency and high-frequency information for image deblurring. To better explore the spatially-variant structures, we develop a learnable low-pass filter based on the self-attention mechanism. Furthermore, we propose a wavelet-based feature fusion method to effectively fuse the features with different spatial resolutions. By formulating the proposed method into an end-to-end trainable network, we show that it performs favorably against state-of-the-art methods on benchmarks.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China under Grant 2022ZD0118801, the National Natural Science Foundation of China under Grants 62272233, 61922043, 61925204, and U22B2049, and the Fundamental Research Funds for the Central Universities under Grant 30922010910 and 30920041109.

References

[1] Ayan Chakrabarti. A neural approach to blind motion deblurring. In *ECCV*, pages 221–235, 2016. 2

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 3, 5, 6

[3] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *ICCV*, pages 4641–4650, 2021. 2, 3, 4, 5, 6, 7, 8

[4] Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Revisiting global statistics aggregation for improving image restoration. In *ECCV*, 2022. 5, 6

[5] Hongyun Gao, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Dynamic scene deblurring with parameter selective sharing and nested skip connections. In *CVPR*, pages 3848–3856, 2019. 2, 3, 5

[6] Dong Gong, Jie Yang, Lingqiao Liu, Yanning Zhang, Ian D. Reid, Chunhua Shen, Anton van den Hengel, and Qinfeng Shi. From motion blur to motion flow: A deep learning solution for removing heterogeneous motion blur. In *CVPR*, pages 3806–3815, 2017. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[8] Jun-Jie Huang and Pier Luigi Dragotti. Winnet: Wavelet-inspired invertible network for image denoising. *IEEE TIP*, 31:4377–4392, 2022. 4

[9] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *CVPR*, pages 1646–1654, 2016. 4

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[11] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR*, pages 233–240, 2011. 2

[12] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, pages 8183–8192, 2018. 2, 6

[13] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, pages 8877–8886, 2019. 6, 7

[14] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In *ICIP*, volume 1, pages I–I, 2002. 5

[15] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *CoRR*, abs/2111.11745, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[16] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *ECCV*, pages 783–798, 2014. 2

[17] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 257–265, 2017. 2, 3, 5, 6, 7

[18] Thekke Madam Nimisha, Akash Kumar Singh, and A. N. Rajagopalan. Blur-invariant deep learning for blind-deblurring. In *ICCV*, pages 4762–4770, 2017. 2

[19] Jinshan Pan, Zhe Hu, Zhixun Su, and Ming-Hsuan Yang. Deblurring text images via l0-regularized intensity and gradient prior. In *CVPR*, pages 2901–2908, 2014. 2

[20] Jinshan Pan, Sifei Liu, Deqing Sun, Jiawei Zhang, Yang Liu, Jimmy S. J. Ren, Zechao Li, Jinhui Tang, Huchuan Lu, Yu-Wing Tai, and Ming-Hsuan Yang. Learning dual convolutional neural networks for low-level vision. In *CVPR*, pages 3070–3079, 2018. 3, 4

[21] Jinshan Pan, Deqing Sun, Hanspeter Pfister, and Ming-Hsuan Yang. Blind image deblurring using dark channel prior. In *CVPR*, pages 1628–1636, 2016. 2

[22] Namuk Park and Songkuk Kim. How do vision transformers work? In *ICLR*, 2022. 1

[23] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, pages 184–201, 2020. 6, 7

[24] Christian J. Schuler, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf. Learning to deblur. *IEEE TPAMI*, 38(7):1439–1451, 2016. 2

[25] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, pages 5571–5580, 2019. 6, 7

[26] Maitreya Suin, Kuldeep Purohit, and A. N. Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *CVPR*, pages 3603–3612, 2020. 2, 5, 6

[27] Jian Sun, Wenfei Cao, Zongben Xu, and Jean Ponce. Learning a convolutional neural network for non-uniform motion blur removal. In *CVPR*, pages 769–777, 2015. 2

[28] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 2, 3, 5, 6, 7

[29] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *ECCV*, 2022. 3, 5, 6, 7, 8

[30] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MAXIM: multi-axis MLP for image processing. In *CVPR*, pages 5759–5770, 2022. 3, 5

[31] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17662–17672, 2022. 3, 5

[32] Phillip Ian Wilson and John Fernandez. Facial feature detection using haar classifiers. *Journal of computing sciences in colleges*, 21(4):127–133, 2006. 5

[33] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *ECCV*, pages 126–144, 2020. 5

[34] Fengze Liu Qingli Li Wei Shen Xintian Mao, Yiming Liu and Yan Wang. Intriguing findings of frequency selection for image deblurring. In *AAAI*, 2023. 1

[35] Li Xu, Shicheng Zheng, and Jiaya Jia. Unnatural L0 sparse representation for natural image deblurring. In *CVPR*, pages 1107–1114, 2013. 2

[36] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729, 2022. 2, 3, 4, 5, 6, 8

[37] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 2, 3, 5, 6, 8

[38] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, pages 5978–5986, 2019. 2, 3, 5

[39] Jiawei Zhang, Jinshan Pan, Jimmy S. J. Ren, Yibing Song, Linchao Bao, Rynson W. H. Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, pages 2521–2529, 2018. 2

[40] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *IJCV*, 130(9):2103–2130, 2022. 1