# Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring

Lingshun Kong[1], Jiangxin Dong[1*], Jianjun Ge[2], Mingqiang Li[2], and Jinshan Pan[1*]
[1]School of Computer Science and Engineering, Nanjing University of Science and Technology
[2]Information Science Academy, China Electronics Technology Group Corporation

## Abstract

*We present an effective and efficient method that explores the properties of Transformers in the frequency domain for high-quality image deblurring. Our method is motivated by the convolution theorem that the correlation or convolution of two signals in the spatial domain is equivalent to an element-wise product of them in the frequency domain. This inspires us to develop an efficient frequency domain-based self-attention solver (FSAS) to estimate the scaled dot-product attention by an element-wise product operation instead of the matrix multiplication in the spatial domain. In addition, we note that simply using the naive feed-forward network (FFN) in Transformers does not generate good deblurred results. To overcome this problem, we propose a simple yet effective discriminative frequency domain-based FFN (DFFN), where we introduce a gated mechanism in the FFN based on the Joint Photographic Experts Group (JPEG) compression algorithm to discriminatively determine which low- and high-frequency information of the features should be preserved for latent clear image restoration. We formulate the proposed FSAS and DFFN into an asymmetrical network based on an encoder and decoder architecture, where the FSAS is only used in the decoder module for better image deblurring. Experimental results show that the proposed method performs favorably against the state-of-the-art approaches.*

## 1. Introduction

Image deblurring aims to restore high-quality images from blurred ones. This problem has achieved significant progress due to the development of various effective deep models with large-scale training datasets.

Most state-of-the-art methods for image deblurring are mainly based on deep convolutional neural networks (CNNs). The main success of these methods is due to developing kinds of network architectural designs, for example, the multi-scale [4, 16, 22] or multi-stage [31, 32] network architectures, generative adversarial learning [11, 12], physics model inspired network structures [6–8, 17, 33], and so on. As the basic operation in these networks, the convolution operation is a spatially-invariant local operation, which does not model the spatially variant properties of the image contents. Most of them use larger and deeper models to remedy the limitation of the convolution. However, simply increasing the capacity of deep models does not always lead to better performance as shown in [17, 33].

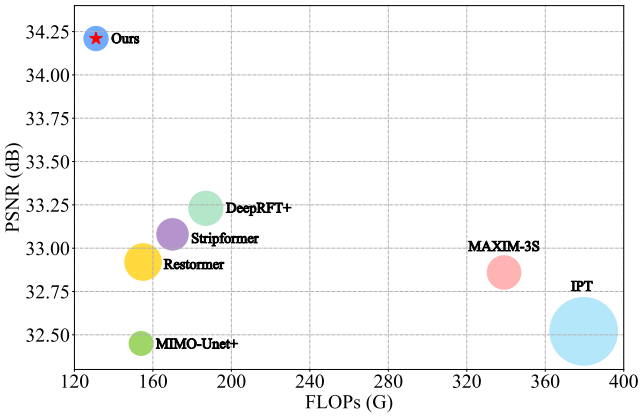



Figure 1. Comparisons of the proposed method and state-of-the-art ones on the GoPro dataset [16] in terms of accuracy, floating point operations (FLOPs), and network parameters. The circle size indicates the number of the network parameter.

Different from the convolution operation that models the local connectivity, Transformers are able to model the global contexts by computing the correlations of one token to all other tokens. They have been shown to be an effective approach in lots of high-level vision tasks and also have great potential to be the alternatives of deep CNN models. In image deblurring, the methods based on Transformers [27, 30] also achieve better performance than the CNN-based methods. However, the computation of the scaled dot-product attention in Transformers leads to quadratic space and time complexity in terms of the number of tokens. Although using smaller and fewer tokens can reduce the space and

*Corresponding author: Jiangxin Dong and Jinshan Pan.

time complexity, such strategy cannot model the long-range information of features well and usually leads to significant artifacts when handling high-resolution images, which thus limits the performance improvement.

To alleviate this problem, most approaches use the downsampling strategy to reduce the spatial resolution of features [26]. However, reducing the spatial resolution of features will cause information loss and thus affect the image deblurring. Several methods reduce the computational cost by computing the scaled dot-product attention in terms of the number of features [29, 30]. Although the computational cost is reduced, the spatial information is well not explored, which may affect the deblurring performance.

In this paper, we develop an effective and efficient method that explores the properties of Transformers for high-quality image deblurring. We note that the scaled dot-product attention computation is actually to estimate the correlation of one token from the query and all the tokens from the key. This process can be achieved by a convolution operation when rearranging the permutations of tokens. Based on this observation and the convolution theorem that the convolution in the spatial domain equals a point-wise multiplication in the frequency domain, we develop an efficient frequency domain-based self-attention solver (FSAS) to estimate the scaled dot-product attention by an element-wise product operation instead of the matrix multiplication. Therefore, the space and time complexity can be reduced to $O(N)$ $O(N \log N)$ for each feature channel, where $N$ is the number of the pixels.

In addition, we note that simply using the feed-forward network (FFN) by [30] does not generate good deblurred results. To generate better features for latent clear image restoration, we develop a simple yet effective discriminative frequency domain-based FFN (DFFN). Our DFFN is motivated by the Joint Photographic Experts Group (JPEG) compression algorithm. It introduces a gated mechanism in the FFN to discriminatively determine which low- and high-frequency information should be preserved for latent clear image restoration.

We formulate the proposed FSAS and DFFN into an end-to-end trainable network based on an encoder and decoder architecture to solve image deblurring. However, we find that as features of shallow layers usually contain blur effects, applying the scaled dot-product attention to shallow features does not effectively explore global clear contents. As the features from deep layers are usually clearer than those from shallow layers, we develop an asymmetric network architecture, where the FSAS is only used in the decoder module for better image deblurring. We analyze that the exploring properties of Transformers in the frequency domain is able to facilitate blur removal. Experimental results demonstrate that the proposed method generates favorable results against state-of-the-art methods in terms of accuracy and efficiency (Figure 1).

The main contributions are summarized as follows:

- We develop an efficient frequency domain-based self-attention solver to estimate the scaled dot-product attention. Our analysis demonstrates that using the frequency domain-based solver reduces the space and time complexity and is much more effective and efficient.
- We propose a simple yet effective discriminative frequency domain-based FFN based on the JPEG compression algorithm to discriminatively determine which low and high-frequency information should be preserved for latent clear image restoration.
- We develop an asymmetric network architecture based on an encoder and decoder network, where the frequency domain-based self-attention solver is only used in the decoder module for better image deblurring.
- We analyze that the exploring properties of Transformers in the frequency domain is able to facilitate blur removal and show that our approach performs favorably against state-of-the-art methods.

## 2. Related Work

**Deep CNN-based Image deblurring methods.** In recent years, we have witnessed significant advances in image deblurring due to the development of different deep CNN models [3, 4, 9, 16, 22, 31, 32]. In [16], Nah et al. propose a deep CNN based on a multi-scale framework to directly estimate clear images from blurred ones. To better utilize the information of each scale in multi-scale framework, Tao et al. [22] develop an effective scale recurrent network. Gao et al. [9] propose a selective network parameter sharing method to improve [16, 22].

As using more scales does not improve the performance significantly, Zhang et al. [32] develop an effective network based on multi-patch strategy. The deblurring process is achieved stage by stage. To better explore the features from different stages, Zamir et al. [31] propose a cross-stage feature fusion for better performance. In order to reduce the computational cost of the methods based on multi-scale framework, Cho et al. [4] present a multi-input and multi-output network. Chen et al. [3] analyze the baseline modules and simplify them for better image restoration. As demonstrated in [30], the convolution operation is spatial invariant and does not effectively model the global contexts for image deblurring.

**Transformers and their applications to image deblurring.** As the Transformer [25] can model the global contexts and achieves significant progress in lots of high-level vision tasks (e.g., image classification [14], object detection [1, 34] and semantic segmentation [28, 35]), it has been developed to solve image super-resolution [13], image deblurring [24, 30] and image denoise [2, 27]. To reduce the computational cost of Transformer, Zamir et al. [30] propose an efficient Transformer model by computing the scaled dot-product attention

in the feature depth domain. This method can effectively explore information from different features along the channel dimension. However, the spatial information that is vital for image restoration is not fully explored. Tsai et al. [24] simplify the calculation of self-attention by constructing intra and inter strip tokens to replace the global attention. Wang et al. [27] propose a Transformer based on a UNet which uses non-overlapping window-based self-attention for single image deblurring. Although using the splitting strategy reduces the computational cost, the coarse splitting does not fully explore the information of each patch. Moreover, the scaled dot-product attention in these methods usually needs the complex matrix multiplication whose the space and time complexity is quadratic.

Different from these methods, we develop an efficient Transformer-based method that explores the property of the frequency domain to avoid the complex matrix multiplication for the scaled dot-product attention.

## 3. Proposed Method

Our goal is to present an effective and efficient method to explore the properties of Transformers for high-quality image deblurring. To this end, we first develop an efficient frequency domain-based self-attention solver to estimate the scaled dot-product attention. To refine the features estimated by the frequency domain-based solver, we further develop a discriminative frequency domain-based feed-forward network. We formulate these above approaches into an end-to-end trainable network based on an encoder and decoder architecture to solve image deblurring, where the frequency domain-based self-attention solver for the estimation of the scaled dot-product attention is used in the decoder module for better feature representation. Figure 2(a) shows the overview of the proposed method. In the following, we present the details of each component.

### 3.1. Frequency domain-based self-attention solver

Given the input feature $X$ with a spatial resolution of $H \times W$ pixels and $C$ channels, existing vision Transformers usually first compute the features $F_q$, $F_k$, and $F_v$ by applying linear transformations $W_q$, $W_k$, and $W_v$ to $X$. Then, they apply the unfolding function to the features $F_q$, $F_k$, and $F_v$ to extract image patches $\{q_i\}_{i=1}^{N}$, $\{k_i\}_{i=1}^{N}$, and $\{v_i\}_{i=1}^{N}$, where $N$ denotes the number of extracted patches. By applying a reshape operation to the extracted patches, the query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ can be obtained by:

$$\mathbf{Q} = \mathcal{R}(\{q_i\}_{i=1}^{N}),\ \mathbf{K} = \mathcal{R}(\{k_i\}_{i=1}^{N}),\ \mathbf{V} = \mathcal{R}(\{v_i\}_{i=1}^{N}), \tag{1}$$

where $\mathcal{R}$ denotes the reshape function which ensures that $\{\mathbf{K}, \mathbf{Q}, \mathbf{V}\} \in R^{N \times (CH_pW_p)}$, $H_p$ and $W_p$ denote the height and width of extracted patches. Based on the obtained query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$, the scaled dot-product attention is achieved by:

$$V_{att} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{CH_pW_p}}\right) V. \tag{2}$$

The attention map computation involves the matrix multiplication of $\mathbf{Q}\mathbf{K}^\top$ whose space complexity and time complexity are $O(N^2)$ and $O(N^2C)$. It is not affordable if the image resolution and the number of the extracted patches are large. Although using downsampling operation to reduce the image resolution or non-overlapping method to extract fewer patches will alleviate the problem, these strategies would lead to information loss and limit the ability to model details within and across each patch [29].

We note that each element of $\mathbf{Q}\mathbf{K}^\top$ is obtained by the inner product:

$$\left(\mathbf{Q}\mathbf{K}^\top\right)_{ij} = \langle \mathbf{q}_i, \mathbf{k}_j \rangle, \tag{3}$$

where $\mathbf{q}_i$ and $\mathbf{k}_j$ are the vectorized forms of $i$-th and $j$-th patches from $F_q$ and $F_k$. Based on (3), if we apply reshape functions to $\mathbf{q}_i$ and all the patches $\mathbf{k}_j$, respectively, all the i-th column elements of $\mathbf{Q}\mathbf{K}^\top$ can be obtained by a convolution operation, i.e., $\widetilde{q}_i \otimes \widetilde{K}$, where $\widetilde{q}_i$ and $\widetilde{K}$ denote the reshaped results of $\mathbf{q}_i$ and $\mathbf{k}_j$; $\otimes$ denotes the convolution operation.

According to the convolution theorem, the correlation or convolution of two signals in the spatial domain is equivalent to an element-wise product of them in the frequency domain. Therefore, *a natural question is that can we efficiently estimate the attention map by an element-wise product operation in a frequency domain instead of computing the matrix multiplication of* $\mathbf{Q}\mathbf{K}^\top$ *in the spatial domain?*

To this end, we develop an effective frequency domain-based self-attention solver. Specifically, we first obtain $F_q$, $F_k$, and $F_v$ by a $1 \times 1$ point-wise convolution and $3 \times 3$ depth-wise convolution. Then, we apply the fast Fourier transform (FFT) to the estimated features $F_q$ and $F_k$ and estimate the correlation of $F_q$ and $F_k$ in the frequency domain by:

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

where $\mathcal{F}(\cdot)$ denotes the FFT, $\mathcal{F}^{-1}(\cdot)$ denotes the inverse FFT, and $\overline{\mathcal{F}(\cdot)}$ denotes the conjugate transpose operation. Finally, we estimate the aggregated feature by:

$$V_{att} = \mathcal{L}(A)\, F_v, \tag{5}$$

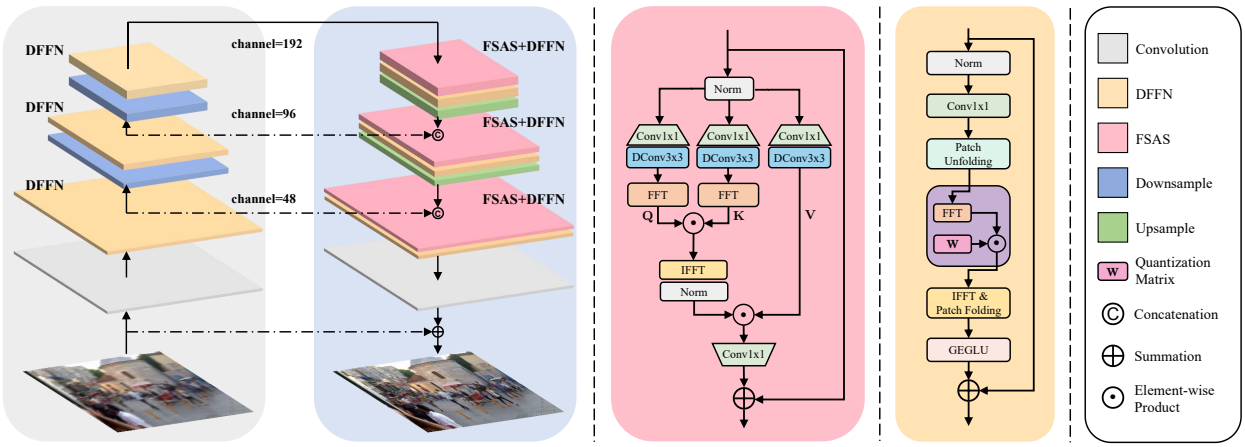
where a layer norm $\mathcal{L}(\cdot)$ is used to normalize $A$. Finally, we generate the output feature of FSAS by:

$$X_{att} = X + \text{Conv}_{1\times1}(V_{att}), \tag{6}$$

where $\text{Conv}_{1\times1}(\cdot)$ denotes a convolution with filter size of $1 \times 1$ pixel. The detailed network architecture of the proposed FSAS is shown in Figure 2(b).

### 3.2. Discriminative frequency domain-based FFN

The FFN is used to improve the features by the scaled dot-product attention. Thus, it is important to develop an




(a) (b) (c)

Figure 2. Network architectures. (a) The proposed asymmetric encoder-decoder network that only contains DFFN in the encoder module and both FSAS and DFFN in the decoder module for image deblurring. (b) The proposed FSAS module. (c) The proposed DFFN module.

effective FFN to generate the features that facilitate the latent clear image reconstruction. As not all the low-frequency information and high-frequency information help latent clear image restoration, we develop a DFFN that can adaptively determine which frequency information should be preserved. However, how to effectively determine which frequency information is important. Motivated by the JPEG compression algorithm, we introduce a learnable quantization matrix $\mathbf{W}$ and learn it by an inverse method of JPEG compression to determine which frequency information should be preserved. The proposed DFFN can be formulated by:

$$\begin{aligned} X_1 &= \mathrm{Conv}_{1\times1}(\mathcal{L}(X_{att})) \\ X_1^f &= \mathcal{F}(\mathcal{P}(X_1)) \\ X_2 &= \mathcal{F}^{-1}(\mathbf{W}X_1^f) \\ X_{out} &= \mathcal{G}\left(\mathcal{P}^{-1}(X_2)\right) + X_{att}, \end{aligned} \tag{7}$$

where $\mathcal{P}(\cdot)$ and $\mathcal{P}^{-1}(\cdot)$ denote the patch unfolding and folding operations in the JPEG compression method; $\mathcal{G}$ denotes GEGLU function by [19]. The detailed network architecture of the proposed DFFN is shown in Figure 2(c).

### 3.3. Asymmetric encoder-decoder network

We embed the proposed FSAS and DFFN into a network based on an encoder and decoder architecture. We note that most existing methods usually use symmetric architectures in the encoder and decoder modules. For example, if the FSAS and DFFN are used in the encoder module, they are also used in the decoder module. We note that the features extracted by encoder module are shallow ones, which usually contain blur effects compared to the deep features from the decoder module. However, the blur usually changes similarity of two similar patches from clear features. Thus, using the FSAS in the encoder module may not estimate the similarity correctly, which accordingly affects image restoration. To overcome this problem, we embed the FSAS into the decoder module, which leads to an asymmetric architecture for better image deblurring. Figure 2(a) shows the network architecture of the proposed asymmetric encoder-decoder network.

Finally, given a blurred image $B$, the restored image $I$ is estimated by the asymmetric encoder-decoder network:

$$I = \mathcal{N}(B) + B, \tag{8}$$

where $\mathcal{N}$ denotes the asymmetric encoder-decoder network.

## 4. Experimental Results

In this section, we evaluate our method and compare it with state-of-the-art ones using public benchmark datasets.

### 4.1. Datasets and parameter settings

**Datasets.** We evaluate our method on commonly used image deblurring datasets including the GoPro dataset [16], the HIDE dataset [20], and the RealBlur dataset [18]. We follow the protocols of existing methods for fair comparisons.

**Parameter settings.** We use the same loss function as [4] to constrain the network and train it using the Adam [10] optimizer with default parameters. The initial value of the learning rate is $10^{-3}$ and is updated with the cosine annealing strategy after 600,000 iterations. The minimum value of the learning rate is $10^{-7}$. The patch size is empirically set to be $256 \times 256$ pixels and the batch size is set to be 16. We adopt the same data augmentation method as [30] during the training. The patch size for the quantization matrix estimation is empirically set to be $8 \times 8$ based on the JPEG compression method. In the implementation, the tensor form of the quantization matrix is set as $[\texttt{batch\_sze}, \texttt{channel\_num}, 8, 8]$, where `batch_size` and `channel_num` are the number of batches and features. The quantization matrix is jointly learned with other parameters by solving loss functions during the training. Similarity, we also use the patch size of $8\times8$ pixels when computing the self-attention (4). Due to the page limit, we include more experimental results in the supplemental material. The training code and models are available at https://github.com/kkkls/FFTformer.

(a) Blurred image (b) GT (c) MIMO-Unet+ [4] (d) Restormer [30]

(e) Stripformer [24] (f) Restormer-local [5] (g) NAFNet [3] (h) Ours

Figure 3. Deblurred results on the GoPro dataset [16]. The deblurred results in (c)-(g) still contain significant blur effects. The proposed method generates a clearer image. For example, the characters and boundaries are much clearer.

Table 1. Quantitative evaluations on the GoPro dataset [16]. The average runtime is tested on images with size of $256 \times 256$ pixels.

| Methods | PSNRs | SSIMs | Parameters (M) | Avg. runtime |
|---|---|---|---|---|
| DeblurGAN-v2 [12] | 29.55 | 0.9340 | 60.9 | 0.04s |
| SRN [22] | 30.26 | 0.9342 | 6.8 | 0.07s |
| DMPHN [32] | 31.20 | 0.9453 | 21.7 | 0.21s |
| SAPHN [21] | 31.85 | 0.9480 | 23.0 | - |
| MIMO-Unet+ [4] | 32.45 | 0.9567 | 16.1 | 0.02s |
| MPRNet [31] | 32.66 | 0.9589 | 20.1 | 0.09s |
| IPT [2] | - | - | 114 | 0.50s |
| DeepRFT+ [15] | 33.23 | 0.9632 | 23.0 | 0.09s |
| Restormer [30] | 32.92 | 0.9611 | 26.1 | 0.08s |
| Uformer-B [27] | 33.06 | 0.9670 | 50.9 | 0.07s |
| Stripformer [24] | 33.08 | 0.9624 | 19.7 | 0.04s |
| MPRNet-local [5] | 33.31 | 0.9637 | 20.1 | 0.11s |
| Restormer-local [5] | 33.57 | 0.9656 | 26.1 | 0.42s |
| NAFNet [3] | 33.71 | 0.9668 | 67.9 | 0.04s |
| Ours | **34.21** | **0.9692** | 16.6 | 0.13s |

## 4.2. Comparisons with the state of the arts

We compare our method with state-of-the-art ones and use the PSNR and SSIM to evaluate the quality of restored images.

**Evaluations on the GoPro dataset.** We first evaluate our method on the commonly used GoPro dataset by [16]. For fair comparisons, we follow the protocols of this dataset and retrain or fine-tune the deep learning methods that are not trained on this dataset. Table 1 shows the quantitative evaluation results. Our method generates the results with the highest PSNR and SSIM values. Compared to the state-of-the-art CNN-based methods, NAFNet [3], the PSNR gain of our method is at least 0.5dB higher than NAFNet, while the number of the proposed model parameters is a quarter of the NAFNet. In addition, compared to the Transformer-based methods [24,27,30], our method has fewer model parameters while the performance is better.

As the FFT implementation is not optimized well in PyTorch, using FFTs in PyTorch needs more runtime. However, the runtime of our method is still competitive against state-of-art methods. Moreover, our method is at least $4\times$ faster than IPT [2] that is based on the original self-attentions.

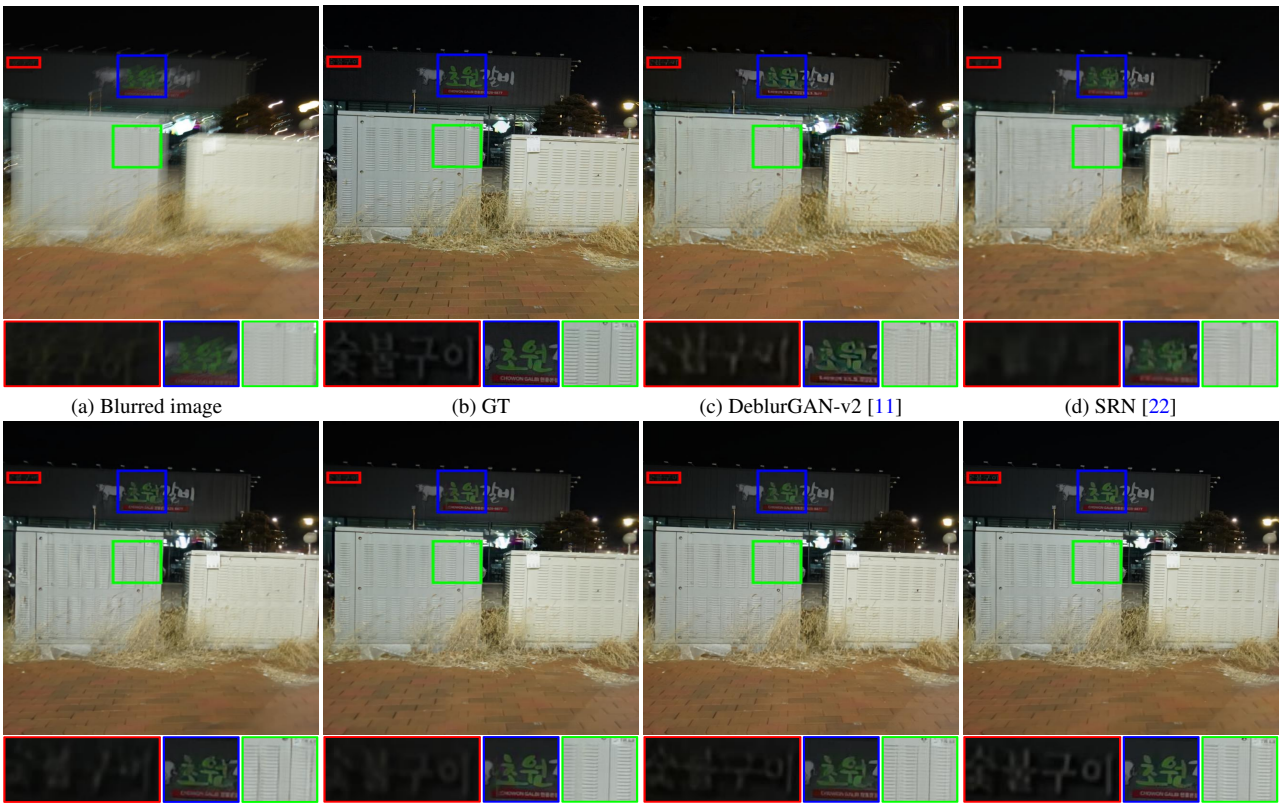
Table 2. Quantitative evaluations on the RealBlur dataset [18] in terms of PSNR and SSIM.

| Methods | Realblur-R | | Realblur-J | |
|---|---|---|---|---|
| | PSNRs | SSIMs | PSNRs | SSIMs |
| DeblurGAN-v2 [12] | 36.44 | 0.9347 | 29.69 | 0.8703 |
| SRN [22] | 38.65 | 0.9652 | 31.38 | 0.9091 |
| MIMO-Unet+ [4] | - | - | 31.92 | 0.9190 |
| BANet [23] | 39.55 | 0.9710 | 32.00 | 0.9230 |
| DeepRFT+ [15] | 39.84 | 0.9721 | 32.19 | 0.9305 |
| Stripformer [24] | 39.84 | **0.9737** | 32.48 | 0.9290 |
| Ours | **40.11** | 0.9732 | **32.62** | **0.9326** |

Figure 3 shows visual comparisons of the proposed method and the evaluated ones on the GoPro dataset. As demonstrated by [30], the CNN-based methods [3,4] do not effectively explore non-local information for latent clear image restoration. Therefore, the deblurred results by the methods [3,4] still contain significant blur effect as shown in Figure 3(c) and (g). The Transformer-based methods [5,24,30] are able to model the global contexts for image deblurring. However, some main structures, e.g., characters and chairs, are not recovered well (see Figure 3(d)-(f)).

In contrast to existing Transformer-based methods that are based on the spatial domain, we develop an efficient frequency domain-based Transformer, where the proposed DFFN is able to discriminately estimate useful frequency information for latent clear image restoration. Thus, the deblurred results contain clear structures, and the characters are much clearer as shown in Figure 3(h).

**Evaluations on the RealBlur dataset.** We further evaluate



(a) Blurred image (b) GT (c) DeblurGAN-v2 [11] (d) SRN [22]

(e) MIMO-Unet+ [4] (f) DeepRFT+ [15] (g) Stripformer [24] (h) Ours

Figure 4. Deblurred results on the RealBlur dataset [18]. The characters or the structural details in (c)-(g) are not recovered well. The proposed method generates an image with much clearer characters and structural details.

Table 3. Quantitative evaluations on the HIDE dataset [20]. We use the models trained on the GoPro dataset [16] for fair comparisons.

| Methods | PSNRs | SSIMs | Parameters (M) |
|---|---|---|---|
| DeblurGAN-v2 [12] | 26.61 | 0.8750 | 60.9 |
| SRN [22] | 28.36 | 0.9040 | 6.8 |
| DMPHN [32] | 29.09 | 0.9240 | 21.7 |
| SAPHN [21] | 29.98 | 0.9300 | 23.0 |
| MIMO-Unet+ [4] | 29.99 | 0.9304 | 16.1 |
| MPRNet [31] | 30.96 | 0.9397 | 20.1 |
| Stripformer [24] | 31.03 | 0.9395 | 19.7 |
| MPRNet-local [5] | 31.19 | 0.9418 | 20.1 |
| Restormer [30] | 31.22 | 0.9423 | 26.1 |
| NAFNet [3] | 31.31 | 0.9427 | 67.9 |
| Restormer-local [5] | 31.49 | 0.9447 | 26.1 |
| Ours | **31.62** | **0.9455** | 16.6 |

our method on the RealBlur dataset by [18] and follow the protocols of this dataset for fair comparisons. The test dataset of [18] includes a RealBlur-R test set from the raw images and RealBlur-J test set from the JPEG images. Table 2 summarizes the quantitative evaluation results on the above mentioned test sets. The proposed method generates the results with higher PSNR values.

Figure 4 shows the visual comparisons on the RealBlur dataset, where our method generates the results with clearer characters and finer structural details (Figure 4(h)).

**Evaluations on the HIDE dataset.** We then evaluate our method on the HIDE dataset [20], which mainly contains humans. Similar to state-of-the-art methods [4, 31], we directly use the models of the evaluated methods, which are trained on the GoPro dataset for test. Table 3 shows that the quality of the deblurred images generated by the proposed method is better than the evaluated methods, suggesting that our method has a better generalization ability as models are not trained on this dataset.

We show some visual comparisons in Figure 5. We note that the evaluated methods do not recover the humans well. In contrast, our method generates better images. For example, the faces and zipper of clothes are much clearer.

## 5. Analysis and Discussion

We have shown that exploring the properties of Transformers in the frequency domain generates favorable results against state-of-the-art methods. In this section, we provide deeper analysis on the proposed method and demonstrate the effect of the main components. For the ablation studies in this section, we train our method and all the baselines on the GoPro dataset using the batch size of 8 to illustrate the effect of each component in our method.

**Effect of FSAS.** The proposed FSAS is used to reduce the computational cost. According to the properties of FFT, the space and time complexity of the FSAS are $O(N)$ and

(a) Blurred image (b) GT (c) MPRNet [31] (d) Restormer [30]

(e) Stripformer [24] (f) NAFNet [3] (g) Restormer-local [5] (h) Ours

Figure 5. Deblurred results on the HIDE dataset [20]. The deblurred results in (c)-(g) still contain significant blur effects. The proposed method generates much clearer images.

Table 4. Memory and running time comparisons of the Transformer based on our method and the window methods [14, 27]. The size of the test image is $1280 \times 720$ pixels. The test environment is based on a machine with an NVIDIA GeForce RTX 3090 GPU. "GPU memory" denotes the maximum GPU memory consumption that is computed by the "torch.cuda.max_memory_allocated()" function.

| | Window-based method [27] | | Ours | |
|---|---|---|---|---|
| Window size | Avg. runtime | GPU memory | Avg. runtime | GPU memory |
| $8 \times 8$ | 53ms | 6.3G | 44ms | 6.5G |
| $16 \times 16$ | 56ms | 7.1G | 44ms | 6.2G |
| $32 \times 32$ | 89ms | 12.0G | 43ms | 6.0G |
| $64 \times 64$ | - | Out of memory | 42ms | 5.9G |
| $1280 \times 720$ | - | Out of memory | 42ms | 5.9G |

$O(NC \log N)$, which are much lower than $O(N^2)$ and $O(N^2C)$ in the original computation of the scaled dot-product attention, where $C$ is the number of features. We further examine the space and time complexity of the FSAS and the window-based strategy [14, 27] for Transformers. Although the tensors in the frequency domain contain real part and imaginary part, the memory consumption does not increase as we only store half parts of the tensors by FFT based on the conjugate symmetry of FFT in our implementations. Table 4 shows that using the proposed FSAS needs a small GPU memory and is much more efficient compared to the window-based strategy [27]. In addition, as the memory of the FSAS is independent of window size, the memory usage of the FSAS does not increase when the window size becomes larger.

Moreover, as the proposed FSAS is performed in the frequency domain, one may wonder whether the scaled dot-product attention estimated in the spatial domain performs better or not. To answer this question, we compare the F-SAS with the baseline method that performs in the spatial domain (SA w/ SD for short). As the space complexity of the original scaled dot-product attention is $O(N^2)$, it is not affordable to train "SA w/ SD" when using the same settings as the proposed FSAS. We use the Swin Transformer [14] for comparison as it is much more efficient. Table 5 shows the quantitative evaluation results on the GoPro dataset. The method that computes the scaled dot-product attention in the spatial domain does not generate good deblurred results, where its PSNR value is 0.27 lower (see comparisons of "SA w/ SD" and "FSAS+DFFN" in Table 5). The main reason is that although using the shifted window partitioning method reduces the computational cost, it does not fully explore the useful information across different windows. In contrast, the space complexity of the proposed FSAS is $O(N)$ and does not need the shifted window partitioning as an approximation, thus leading to better deblurred results. Figure 6(b) further shows that using the shifted window partitioning method as an approximation of scaled dot-product attention in the spatial domain does not remove blur effectively. In contrast, the proposed FSAS generates clearer images.

Table 5. Quantitative evaluations of each component in the proposed method on the GoPro dataset [16].

| | FSAS | Swin attention | FFN | DFFN | PSNRs/SSIMs |
|---|---|---|---|---|---|
| w/ only FFN | ✗ | ✗ | ✓ | ✗ | 33.19/0.9626 |
| w/ only DFFN | ✗ | ✗ | ✗ | ✓ | 33.55/0.9651 |
| SA w/ SD | ✗ | ✓ | ✗ | ✓ | 33.46/0.9645 |
| FSAS+FFN | ✓ | ✗ | ✓ | ✗ | 33.61/0.9654 |
| FSAS+DFFN | ✓ | ✗ | ✗ | ✓ | 33.73/0.9663 |

Moreover, compared to the baseline method only using FFN ("w/ only FFN"), using the proposed FSAS in this baseline generates much better results, where the PSNR

(a) Blurred image (b) Spatial domain (c) Frequency domain

Figure 6. Effectiveness of the scaled dot-product attention computation in the spatial and frequency domains. Computing the scaled dot-product attention using the FSAS in the frequency domain generates a clearer image.

(a) Blurred image (b) Ours w/o FSAS (c) Ours

Figure 7. Effectiveness of the proposed FSAS on image deblurring. Using the proposed FSAS generates a clearer image.

(a) Blurred image (b) w/ only FFN (c) w/ only DFFN

Figure 8. Effectiveness of the proposed DFFN on image deblurring.

value is 0.42dB higher (see comparisons of "w/ only FFN" and "FSAS+FFN" in Table 5). The visual comparisons in Figure 7(b) and (c) further demonstrate that using the proposed FSAS facilitates the blur removal well, where the boundaries are recovered well as shown in Figure 7(c).

**Effect of DFFN.** The proposed DFFN is used to discriminatively estimate useful frequency information for latent clear image restoration. To demonstrate its effectiveness on image deblurring, we compare the proposed method with two baselines. For the first baseline, we compare the proposed method only using the DFFN (w/ only DFFN for short) and the proposed method only using the original FFN (w/ only FFN for short). For the second baseline, we compare the proposed method with the one that replaces the DFFN with the original FFN in the proposed method (FSAS+FFN). The comparisons of "w/ only DFFN" and "w/ only FFN" in Table 5 show that using the proposed DFFN generates better results, where the PSNR value is 0.36dB higher.

In addition, the comparisons of "FSAS+FFN" and "FSAS+DFFN" in Table 5 show that using the proposed DFFN further improves the performance.

Figure 8 shows the visualization results by these above mentioned baseline methods. Using the proposed DFFN generates better deblurred images, where the windows are recovered well shown in Figure 8(c).

Table 6. Quantitative evaluations of the asymmetric encoder-decoder network on the GoPro dataset.

| Methods | FSAS in enc&dec | FSAS in dec (Ours) |
|---|---|---|
| PSNRs | 33.56 | **33.73** |
| SSIMs | 0.9653 | **0.9663** |

(a) Blurred image (b) FSAS in enc&dec (c) FSAS in dec (Ours)

Figure 9. Effectiveness of the asymmetric encoder-decoder network on image deblurring.

**Effect of the asymmetric encoder-decoder network.** As demonstrated in Section 3.3, the shallow features extracted by encoder module usually contain blur effects that affect the estimations of FSAS. We thus embed it into the decoder module, which leads to an asymmetric encoder-decoder network for better image deblurring. To examine the effect of this network design, we compare the network that puts the FSAS into both the encoder and decoder modules ("FSAS in enc&dec" in Table 6). Table 6 shows that using the FSAS in the decoder module generates better results, where the PSNR value is at least 0.17dB higher. The visual comparisons in Figure 9(b) and (c) further demonstrate that using the FSAS in the decoder module generates better clear images.

## 6. Conclusion

Motivated by the convolution theorem, we have presented an effective and efficient method that explores the properties of Transformers for high-quality image deblurring. We have developed an efficient frequency domain-based self-attention solver (FSAS) to estimate the scaled dot-product attention by an element-wise product operation instead of the matrix multiplication in the spatial domain, where we show that the spatial complexity and the computational complexity are significantly reduced. We further propose a DFFN to discriminatively determine which low and high frequency information of the features should be preserved for latent clear image restoration. Moreover, we develop an asymmetrical network based on an encoder and decoder architecture, where the FSAS is only used in the decoder module for better image deblurring. By training our method in an end-to-end manner, we show that it performs favorably against the state-of-the-art approaches in terms of accuracy and efficiency.

**Acknowledgements**. This work has been partly supported by the National Key R&D Program of China (No. 2018AAA0102001), the National Natural Science Foundation of China (Nos. U22B2049, 62272233, 61922043, U19B2040), and the Fundamental Research Funds for the Central Universities (No. 30920041109).

## References


[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[2] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, 2021. 2, 5

[3] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, 2022. 2, 5, 6, 7

[4] 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*, 2021. 1, 2, 4, 5, 6

[5] Xiaojie Chu, Liangyu Chen, , Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *ECCV*, 2022. 5, 6, 7

[6] Jiangxin Dong, Jinshan Pan, Jimmy S. Ren, Liang Lin, Jinhui Tang, and Ming-Hsuan Yang. Learning spatially variant linear representation models for joint filtering. *IEEE TPAMI*, 44(11):8355–8370, 2022. 1

[7] Jiangxin Dong, Stefan Roth, and Bernt Schiele. Learning spatially-variant MAP models for non-blind image deblurring. In *CVPR*, 2021. 1

[8] Jiangxin Dong, Stefan Roth, and Bernt Schiele. DWDN: deep wiener deconvolution network for non-blind image deblurring. *IEEE TPAMI*, 44(12):9960–9976, 2022. 1

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

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4

[11] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiri Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, 2018. 1, 6

[12] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, 2019. 1, 5, 6

[13] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV Workshops*, 2021. 2

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 2, 7

[15] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021. 5, 6

[16] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 1, 2, 4, 5, 6, 7

[17] Jinshan Pan, Jiangxin Dong, Yang Liu, Jiawei Zhang, Jimmy S. J. Ren, Jinhui Tang, Yu-Wing Tai, and Ming-Hsuan Yang. Physics-based generative adversarial models for image restoration and beyond. *IEEE TPAMI*, 43(7):2449–2462, 2021. 1

[18] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *ECCV*, 2020. 4, 5, 6

[19] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 4

[20] Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *ICCV*, 2019. 4, 6, 7

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

[22] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, 2018. 1, 2, 5, 6

[23] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Banet: Blur-aware attention networks for dynamic scene deblurring. In *CVPR*, 2021. 5

[24] 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. 2, 3, 5, 6, 7

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[26] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 2

[27] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, 2022. 1, 2, 3, 5, 7

[28] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, 2022. 2

[29] Weijian Xu, Yifan Xu, Tyler A. Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *ICCV*, 2021. 2, 3

[30] 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*, 2022. 1, 2, 4, 5, 6, 7

[31] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, 2021. 1, 2, 5, 6, 7

[32] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, 2019. 1, 2, 5, 6

[33] Jiawei Zhang, Jinshan Pan, Jimmy Ren, Yibing Song, Linchao Bao, Rynson W. H. Lau, and Ming-Hsuan Yang. Dynamic scene deblurring using spatially variant recurrent neural networks. In *CVPR*, 2018. 1

[34] Jing Zhang, Jianwen Xie, Nick Barnes, and Ping Li. Learning generative vision transformer with energy-based latent space for saliency prediction. In *NeurIPS*, 2021. 2

[35] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 2