Dual-Domain Attention for Image Deblurring

Yuning Cui[1*], **Yi Tao**[2*], **Wenqi Ren**[3†], **Alois Knoll**[1]

[1]Technical University of Munich
[2]MIT Universal Village Program
[3]Shenzhen Campus of Sun Yat-sen University
{yuning.cui, knoll}@in.tum.de, yitao@universal-village.org, renwq3@mail.sysu.edu.cn

Abstract

As a long-standing and challenging task, image deblurring aims to reconstruct the latent sharp image from its degraded counterpart. In this study, to bridge the gaps between degraded/sharp image pairs in the spatial and frequency domains simultaneously, we develop the dual-domain attention mechanism for image deblurring. Self-attention is widely used in vision tasks, however, due to the quadratic complexity, it is not applicable to image deblurring with high-resolution images. To alleviate this issue, we propose a novel spatial attention module by implementing self-attention in the style of dynamic group convolution for integrating information from the local region, enhancing the representation learning capability and reducing computational burden. Regarding frequency domain learning, many frequency-based deblurring approaches either treat the spectrum as a whole or decompose frequency components in a complicated manner. In this work, we devise a frequency attention module to compactly decouple the spectrum into distinct frequency parts and accentuate the informative part with extremely lightweight learnable parameters. Finally, we incorporate attention modules into a U-shaped network. Extensive comparisons with prior arts on the common benchmarks show that our model, named **D**ual-**D**omain **A**ttention **Net**work (DDANet), obtains comparable results with a significantly improved inference speed.

Introduction

In this paper, we address the problem of blind motion deblurring whose aim is to recover the sharp image from a degraded observation caused by motion of camera during sensor exposure or objects in the scene (Kundur and Hatzinakos 1996; Campisi and Egiazarian 2017). This problem is encountered in many diverse technical areas, such as remote sensing (Bertero, Boccacci, and De Mol 2021), medical imaging (Michailovich and Adam 2005), photography (Sroubek and Flusser 2005), and self-driving cars (Zhang et al. 2022a). Deblurring also serves as a fundamental preprocessing measure for some downstream computer vision tasks (Sayed and Brostow 2021). Due to the ill-posed property, it has attracted considerable attention from industrial community and academia over the years.

Considering image deblurring as an inverse problem, numerous approaches have been developed to deal with blind deblurring problems and non-blind deblurring problems (Wang and Tao 2014). Non-blind deblurring means that the blue kernel is assumed to be known and the target image can be generated from both the kernel and the observation. The typical methods for this category include Wiener filter (Wiener et al. 1949) and Richardson-Lucy (Richardson 1972; Lucy 1974). By contrast, dating back to the 1970s, blind image deblurring is more practical and challenging where the blur kernel is unknown. Many novel approaches have been put forward to tackle it by estimating the blue kernel and sharp image successively or simultaneously (Bahat, Efrat, and Irani 2017; Aittala and Durand 2018). However, with various additional constraints, these methods are not applicable in the complicated scenarios.

Recently, with the rapid development of deep learning, multifarious methods based on convolutional neural network (CNN) have emerged as a preferable solution compared to above-mentioned methods, due to their advanced mapping capacity from blurry image to the sharp target (Nah, Hyun Kim, and Mu Lee 2017; Suin, Purohit, and Rajagopalan 2020; Chen et al. 2022; Zamir et al. 2021). Despite the improved performance these methods have brought, with the limited receptive filed of convolution, they struggle to integrate the long-range dependencies. More recently, to alleviate this issue, Transformer methods have been tailored for deblurring to further boost the accuracy of reconstruction (Wang et al. 2022; Tsai et al. 2022). The core component of these approaches is self-attention mechanism where each pixel is modulated by a weighted sum of the whole feature map. Despite many efforts to reduce the quadratic complexity of this technique, these solutions still struggle to have efficient implementation compared to CNN-based methods.

In this paper, we propose an efficient spatial attention module (SAM) that combines the merits of group convolution (Krizhevsky, Sutskever, and Hinton 2012) and self-attention (Vaswani et al. 2017). Our SAM is Tanh-normalized and shares weights over the group dimension, utilizing the group convolution mechanism to mimic the operation of self-attention. Compared to self-attention, SAM performs information aggregation within a small region to reduce the computational complexity and substitutes hyperbolic tangent function for Softmax from the inspiration that

*These authors contributed equally.
†Corresponding author.

not all surrounding pixels have the positive effect on the centered pixel. Importantly, our SAM generates integration weights for each two pixels from contextual information instead of from information that is only related to these two pixels. Different to group-wise convolution, the integration weights in our module are shared in each group across channel dimension which results in further efficiency.

On the other hand, most of the recent deep learning-based methods solve deblurring only in the spatial domain without sufficiently utilizing the discrepancies in the frequency domain. Recently, some works have been proposed to reduce the frequency gap between sharp/blurry image pairs (Zou et al. 2021; Mao et al. 2021; Liu et al. 2020; Zhang et al. 2022b). For instance, SDWNet (Zou et al. 2021) introduces wavelet reconstruction module to decouple the features into various frequency subbands by Wavelet transform, which needs additional computational complexity to perform inverse transform. DeepRFT (Mao et al. 2021) treats high-frequency signals with the normal residual branch and deals with omni-frequency using Fourier transform-based branch. However, these two branches share the same input, and as a consequence, the low-frequency information conveyed to residual branch may disturb the learning of high-frequency signal. MSFS-Net (Zhang et al. 2022b) adopts multiple complicated OctConv (Chen et al. 2019) to conduct frequency separation, and down/upsampling operations are performed frequently, introducing extra computation burdens.

In this work, we develop a simple yet effective and efficient frequency attention module to decouple the feature into low- and high-frequency components, and then accentuate the informative ones via the learnable weights. To this end, we leverage average pooling with different kernel sizes, *i.e.,* 3×3 kernel and global kernel, to obtain two low-frequency parts with different receptive fields. After generating opposite high-frequency components by subtracting low-frequency parts from the input, learnable weights are imposed on the resultant diverse frequency signals to achieve recalibration. In this manner, the various frequency subbands are treated individually. Our strategy enjoys several advantages. Firstly, it works without using FFT or wavelet transform and hence no extra inverse transform is needed. Compared to FFT, it does not require any further processing to distinguish different frequency components. Second, it is compact without employing any postprocessing convolution as in (Mao et al. 2021; Zou et al. 2021).

In summary, our main contributions are as follows:

- We develop a dual-domain attention mechanism to boost image deblurring performance by enhancing representation learning in the spatial and frequency domains.
- We propose a spatial attention module (SAM) that inherits the strengths of convolution and self-attention to capture local dependencies efficiently.
- We devise a frequency attention module (FAM) that performs controlled frequency transformation by accentuating the more informative frequency parts with an extremely lightweight implementation.
- We conduct comprehensive comparisons with prior arts to demonstrate the effectiveness of our attention method.

Related Work

Image deblurring. Recently, CNN-based architectures have outperformed the conventional deblurring approaches. As a seminal technique, DeepDeblur (Nah, Hyun Kim, and Mu Lee 2017) directly learns the regression between image pairs, exhibiting superiority in deblurring over kernel-based frameworks. Thereafter, equipped with various advanced functional units and modules, *e.g.,* dilated convolution, UNet and attention mechanism, abundant CNN networks have been investigated to improve performance (Zamir et al. 2021; Chen et al. 2021b; Yuan, Su, and Ma 2020). More recently, some researchers have introduced Transformer into image deblurring to capture long-range dependencies (Zamir et al. 2022; Wang et al. 2022; Chen et al. 2021a). In this paper, we pursue a dual-domain attention mechanism to address deblurring.

Self-attention in image deblurring. Attention mechanisms, especially self-attention (Vaswani et al. 2017), play an important role to attend to relevant information and enhance the representation learning ability in image deblurring frameworks (Purohit and Rajagopalan 2020). For instance, efficient self-attention is leveraged in an encoder-decoder architecture to obtain better representation (Suin, Purohit, and Rajagopalan 2020). Sparse non-local attention module is proposed in (Purohit et al. 2021) to cope with the spatially-varying degradation.

Of late, various methods have been developed in computer vision community to reduce the quadratic complexity of canonical self-attention (SA) (Wu et al. 2019; Li et al. 2022a; Liang et al. 2021; Li et al. 2022b). In the context of image deblurring, SA across channel dimension is proposed in Restormer (Zamir et al. 2022) to reduce complexities. Uformer (Wang et al. 2022) performs SA within non-overlapping local windows as in Swin Transformer (Liu et al. 2021). Stripformer (Tsai et al. 2022) introduces intra- and inter-strip attentions for dynamic scene image deblurring. In this paper, we also focus on SA to achieve spatial attention, but implement in a convolution style to make it more efficient.

Frequency analysis in image deblurring. Fourier analysis is widely used in traditional image deblurring methods owing to convolution theorem (Ayers and Dainty 1988; Delbracio and Sapiro 2015; Keuper et al. 2013). Recently, some CNN-based frameworks have been designed to bridge the frequency gap between blurry/sharp image pairs (Liu et al. 2020; Cui et al. 2023). SDWNet uses wavelet transform to separate the input into four frequency subbands and treats each with individual convolution to avoid interference between different frequency parts (Zou et al. 2021). DeepRFT leverages FFT-based branch to process frequency information (Mao et al. 2021). MSFS-Net adopts multiple OctConv (Chen et al. 2019) operation to decompose frequency (Zhang et al. 2022b). We explore a simpler frequency attention module to first decouple the spectrum with average pooling technique, and then emphasize the useful frequency components and suppress the less informative ones via learnable parameters.

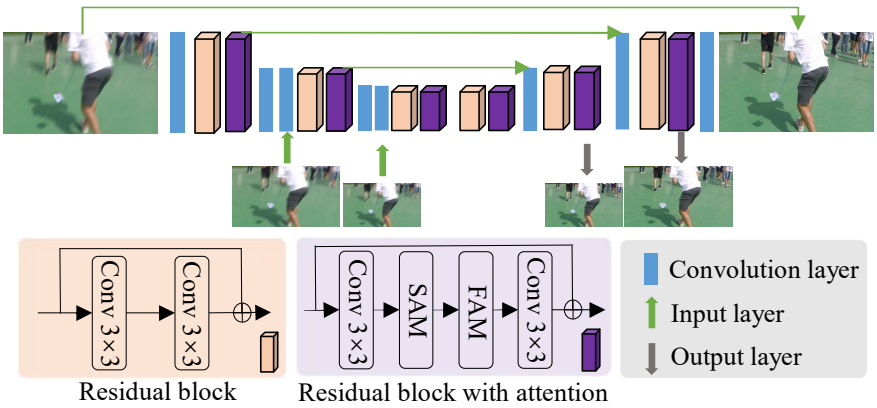

Figure 1: The pipeline of our framework. We insert our spatial and frequency attention modules between two convolution layers. The input and output layers are borrowed from MIMO-UNet (Cho et al. 2021).

Proposed Algorithm

In this section, we first introduce two attention modules in detail. Then we describe the overall architecture of the proposed image deblurring network. The loss functions follow in the final part.

Spatial Attention Module (SAM)

To enhance the representation capacity of the extracted features in spatial domain, the proposed SAM mimics the operation of MHSA in the style of convolution. We first revisit the formulation of MHSA (Vaswani et al. 2017):

$$\mathrm{MHSA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,...,\mathrm{head}_h)W^O \quad (1)$$

$$\mathrm{head}_i = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (2)$$

where, W_i^Q, W_i^K, W_i^V are parameter matrices to project Q, K and V into different representation subspaces. The core component of each head is scaled dot-product attention:

$$\begin{aligned}\mathrm{Attention}(Q,K,V) &= AV \\ &= \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V\end{aligned} \quad (3)$$

Given the input shape of $\mathbb{R}^{H\times W\times C}$, the size of the attention map (A) and the outcome is $\mathbb{R}^{HW\times HW}$ and $\mathbb{R}^{HW\times C}$, respectively. We assume that Q, K, V in Eq. 3 have the same number of channels with the input for convenience. Hence, $Q, K, V \in \mathbb{R}^{HW\times C}$.

Next, we will show that the weights of other pixels to a target pixel are calculated locally. To be specific, for a single resultant pixel of Eq. 3, it is obtained by,

$$y_{i,j} = \sum_{p=1}^{HW} A_{i,p}V_{p,j} \quad (4)$$

where, i, p, j denote the coordinates. The contribution of $V_{i+1,j}$ to the calculation of $y_{i,j}$ is the value of $A_{i,i+1}$, which is generated by,

$$A_{i,i+1} = \sum_{q=1}^{C} Q_{i,q}K^T_{q,i+1} \quad (5)$$

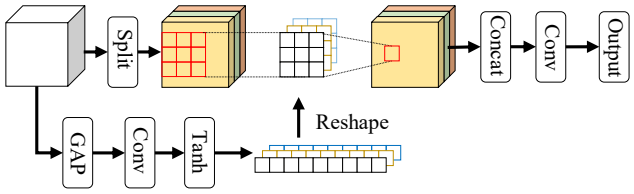

Figure 2: Spatial attention module.

Here, we leave out the scale factor in Eq. 3 for simplicity. From Eq. 5, we can conclude that the weight of another pixel to a target one is generated by only considering information of these two pixels across channel dimension without taking other pixels into account.

To obtain weights for each two pixels from contextual information, our SAM first adopts global average pooling (GAP) to generate a global feature, and then a convolution layer is utilized to adjust the channel dimension. As claimed in (Purohit et al. 2021), propagating all information across spatial domain has a side effect. Hence, we reduce the scope of propagating and conduct local self-attention. In addition, different from the vanilla self-attention, we instead utilize the hyperbolic tangent activation function to generate negative weights for harmful pixels, and as a consequence, the negative effects of these pixels can be suppressed. Formally, given the input $X \in \mathbb{R}^{C\times H\times W}$, our attention weights for all groups are obtained by,

$$W = \mathrm{Tanh}(W_1 * \mathrm{GAP}(X)) \quad (6)$$

where, W_1 is the parameter matrix of convolution layer, and $*$ denotes convolution operation. $W \in \mathbb{R}^{g\times k^2}$, where k is the kernel size of attention, and g is the number of groups.

Following group convolution, we split the input X into several groups, but each group shares the same attention weights across the channel and spatial dimensions. This measure can significantly reduce the number of attention weights and parameters, and hence ease the difficulty of training. After obtaining the attention weights and group feature, the local self-attention is described by,

$$S_i = W_i * X_i \quad (7)$$

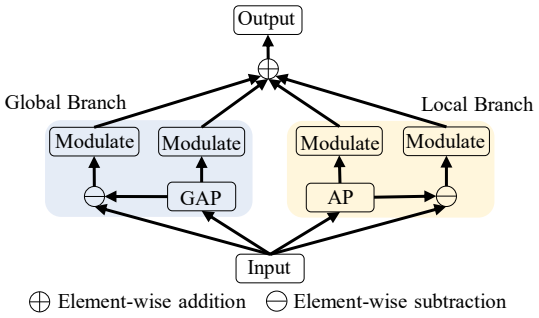

Figure 3: Frequency attention module. GAP denotes the Global Average Pooling, and AP is Average Pooling with the kernel size of 3×3. Modulation is implemented by channel-wise recalibration where the attention weights are set as directly learnable parameters without introducing any additional sublayer.

where, i is the index of group. To promote interactions between different groups as MHSA does in Eq. 1, we apply another convolution layer to get the final output of SAM,

$$\mathrm{SAM}(X) = \mathrm{Concat}(S_1, S_2, ..., S_g)W_2 \tag{8}$$

To summarize, our SAM is different from the vanilla self-attention in following aspects: (*i*) The attention weights are generated based on the contextual information. (*ii*) The scope for integration is limited in a small region to reduce the computational complexity. (*iii*) Negative weights are produced to suppress the effects of harmful pixels via hyperbolic tangent activation function.

Frequency Attention Module (FAM)

(Liu et al. 2020) shows that the low-frequency parts of a sharp image and its blurry counterpart are similar while there are huge discrepancies between high-frequency components. Inspired by this opinion, to enable efficient frequency learning, we design FAM to treat different frequency subbands individually. FAM mainly contains two steps, decomposition and modulation, as illustrated in Figure 3.

Intuitively, the simplest decoupling method is to split the spectrum into two parts, the lowest part and the opposite. To achieve this goal, FAM is established based on average pooling whose result is proportional to the lowest frequency component in the spectrum (Qin et al. 2021).

Given an input $X \in \mathbb{R}^{C\times H\times W}$, we apply global average pooling to extract the lowest frequency component, and the high frequency can be obtained by subtracting the resultant low frequency from X. This process can be formally described as,

$$X_g^l = \mathrm{GAP}(X); \quad X_g^h = X - X_g^l \tag{9}$$

where X_g^l and X_g^h denote the global low- and high-frequency components, respectively.

Furthermore, due to the important role of receptive filed in image deblurring, apart from the global frequency modulation, we also add a fine-grained pooling operation in the local branch (see Figure 3). Specifically, we utilize the average pooling with the kernel size of 3×3 to extract local lowest frequency. Similar to Eq. 9, we can obtain,

$$X_l^l = \mathrm{AP}(X); \quad X_l^h = X - X_l^l \tag{10}$$

where X_l^l and X_l^h are the local low- and high-frequency components. After getting the global low/high frequency and local low/high frequency information in Eq. 9 and Eq. 10, we modulate them via the channel-wise attention weights. The final output of FAM is obtained by adding these elements together,

$$\mathrm{FAM}(X) = W_g^l X_g^l + W_g^h X_g^h + W_l^l X_l^l + W_l^h X_l^h \tag{11}$$

where the channel-wise weights W are the learnable parameters directly optimized via the back-propagation. We do not utilize any additional sublayer to obtain the attention weights for simplicity.

Overall Pipeline

We incorporate two attention modules into an encoder-decoder baseline to establish our dual-domain attention network. Following (Cho et al. 2021), we adopt the multi-input and multi-output mechanisms to ease training difficulty. The bottom part of Figure 1 shows the difference between the altered residual block and the standard one.

On the whole, DDANet is a U-shaped hierarchical network, which has three scales for the encoder and decoder. To be specific, given an observed image, DDANet first applies a single convolution layer to extract the low-level feature. Then the obtained shallow feature is fed into three scales of the encoder. Each scale has 20 residual blocks in total, and we employ FAM/local in the last 4 ones while FAM/global in last 8 blocks. SAM is only used in the last residual block. The number of total blocks are kept identical in each scale. Starting from the very first image, the encoder network progressively reduces the spatial size while doubling the number of channels. The decoder performs in an inverse manner as encoder. In addition, feature-level and image-level skip connections are applied as in (Cho et al. 2021; Mao et al. 2021). Finally, the sharp image is generated by adding the degraded image via the global skip connection.

Loss Function

To facilitate dual-domain learning, we adopt the L_1 loss function in the spatial and frequency domains (Cho et al. 2021).

$$L_{spa} = \sum_{r=1}^{R} \frac{1}{S_r} \|\hat{y_r} - y_r\|_1 \tag{12}$$

$$L_{fre} = \sum_{r=1}^{R} \frac{1}{S_r} \|\mathcal{F}(\hat{y_r}) - \mathcal{F}(y_r)\|_1 \tag{13}$$

$$L = L_{spa} + \lambda L_{fre} \tag{14}$$

where $\hat{y_r}$ is the predicted image, y_r denotes the ground truth, R is the index of multi-scale output. S_r is the normalization factor representing the number of total elements of $\hat{y_r}$. We empirically set the hyper-parameter λ as 0.1 to balance the dual-domain learning.

Method	PSNR↑	SSIM↑	Params	FLOPs
Nah (2017)	29.08	0.914	11.7	-
DeblurGAN (2018)	28.70	0.858	-	-
DeblurGAN-v2 (2019)	29.55	0.934	60.9	-
SRN (2018)	30.26	0.934	6.8	
DBGAN (2020)	31.10	0.942	11.6	759.85
DMPHN (2019)	31.20	0.940	21.7	-
MIMO-UNet+ (2021)	32.45	0.957	16.1	154.41
HINet (2021)	32.71	0.959	88.7	170.5
MPRNet (2021)	32.66	0.959	20.1	777.01
IPT (2021)	32.52	-	114	594
MSFS-Net (2022)	32.73	0.959	-	
KiT (2022)	32.70	0.959	-	
MAXIM-3S (2022)	32.86	0.961	22.2	169.5
Restormer (2022)	32.92	0.961	26.13	140.99
DDANet	**33.07**	**0.962**	16.18	153.51

Table 1: Image deblurring comparisons on the GoPro (Nah, Hyun Kim, and Mu Lee 2017) benchmark dataset.

Experiments

In this section, we first delineate the datasets and the details of implementation. Next, we provide comparisons between DDANet and state-of-the-art deblurring methods to verify the superiority of our network. Finally, ablation studies are presented to validate the effectiveness of each design.

Datasets and Implementation Details

Following recent works (Zamir et al. 2022; Tu et al. 2022), we utilize the GoPro (Nah, Hyun Kim, and Mu Lee 2017) dataset that contains 2,103 blurry/sharp image pairs for training and 1,111 pairs for evaluation. In addition, to validate generalization capability, we directly apply GoPro-trained model to the synthetic dataset (HIDE (Shen et al. 2019)) and real-world dataset (RealBlur (Rim et al. 2020)). PSNR and SSIM (Wang et al. 2004) serve as the metrics.

We train DDANet with Adam (Kingma and Ba 2014) optimizer with the initial learning rate as 1×10^{-4}, which is reduced to 1×10^{-6} via the cosine annealing strategy (Loshchilov and Hutter 2016). The network is trained on 256×256 patches with a batch size of 4 for 3000 epochs, and tested on the full resolution. For data augmentation, horizontal flips are randomly applied with a probability of 0.5. The kernel size of SAM is set as 3×3. Our experiments are performed on an NVIDIA Tesla V100 GPU and Intel Xeon Platinum 8255C CPU. FLOPs are measured on 256×256 patches.

Quantitative and Qualitative Evaluation

Table 1 shows the comparisons with state-of-the-art methods (Zamir et al. 2022; Tu et al. 2022; Lee et al. 2022) on GoPro (Nah, Hyun Kim, and Mu Lee 2017). Our DDANet obtains significantly better accuracy than existing algorithms. To be specific, compared to the advanced Restormer (Zamir et al. 2022), DDANet outperforms it by 0.15 dB with about 10M fewer parameters and comparable computation complexity.

Method	DBGAN	DeepRFT+	MPRNet	Restormer	Ours
PSNR	31.10	32.45	32.66	32.92	33.07
Time (s)	1.447	0.806	1.148	1.218	0.247

Table 2: Inference time comparisons on GoPro test dataset. Inference time is evaluated on 720×1280 images by second. The accuracy of DeepRFT+ (Mao et al. 2021) is obtained by removing the patch-based strategy.

Baseline	SAM	FAM	PSNR	SSIM	Params	FLOPs
✓			31.42	0.948	6.81	67.2
✓	✓		31.82	0.951	6.92	67.6
✓		✓	32.17	0.953	6.82	67.2
✓	✓	✓	32.29	0.958	6.93	67.6

Table 3: Ablation studies for the proposed modules on GoPro (Nah, Hyun Kim, and Mu Lee 2017).

These results demonstrate the better trade-off of our method between accuracy and computational cost.

Moreover, we also test the GoPro-trained model on another synthetic dataset and the real-world dataset without fine-tuning. The results on HIDE (Shen et al. 2019) and RealBlur (Rim et al. 2020) are listed in Table 5 and Table 6, respectively. On HIDE dataset, our method achieves an improvement of 0.65 dB PSNR compared to MIMO-UNet+ (Cho et al. 2021). On the more challenging real-world dataset, DDANet gains 0.03 dB higher score in terms of PSNR than MAXIM-3S (Tu et al. 2022) algorithm, showing the strong generalization capability of our method.

In addition, as shown in Table 2, we evaluate the inference speed of several state-of-the-art deblurring methods on GoPro testset (Nah, Hyun Kim, and Mu Lee 2017) using the released test code and pre-trained models on our equipment. DDANet is almost $5\times$ faster than Restormer (Zamir et al. 2022) with 0.15 dB higher accuracy. Since DeepRFT+ (Mao et al. 2021) shares the similar baseline network with our model, we remove its patch-based testing strategy borrowed from (Zou et al. 2021) for a fair comparison. Despite using DO-Conv (Cao et al. 2022), DeepRFT+ only achieves 32.45 dB PSNR with an inferior inference speed to our DDANet.

Visual comparisons on GoPro (Nah, Hyun Kim, and Mu Lee 2017) are shown in Figure 4. Our approach is more effective in removing motion blurs than other methods.

Ablation Studies

Ablation studies are conducted by applying our designs to the baseline network (Nah, Hyun Kim, and Mu Lee 2017),

Method	None	Sigmoid	Softmax	Tanh
PSNR	31.67	31.67	31.66	31.77

Table 4: SAM with different activation functions. The number of groups is 8 and the convolution in Eq. 8 is not used.

Method	DMPHN	DeblurGAN	DeblurGAN-v2	SRN	DBGAN	MT-RNN	MIMO-UNet+	SPAIR	TTFA	DDANet
PSNR↑	29.09	24.51	26.61	28.36	28.94	29.15	29.99	30.29	30.55	**30.64**
SSIM↑	0.924	0.871	0.875	0.915	0.915	0.918	0.930	0.931	0.935	**0.937**

Table 5: Image deblurring results on HIDE (Shen et al. 2019). Our DDANet is trained only on the GoPro (Nah, Hyun Kim, and Mu Lee 2017) dataset and directly applied to the HIDE benchmark.

Method	DMPHN	DeblurGAN	DeblurGAN-v2	SRN	DBGAN	MT-RNN	MIMO-UNet+	MAXIM-3S	DDANet
PSNR↑	35.70	33.79	35.26	35.66	33.78	35.79	35.54	35.78	**35.81**
SSIM↑	0.948	0.903	0.944	0.947	0.909	**0.951**	0.947	0.947	**0.951**

Table 6: Image deblurring results on RealBlur-R (Rim et al. 2020). Our DDANet is trained only on the GoPro (Nah, Hyun Kim, and Mu Lee 2017) dataset and directly applied to the RealBlur benchmark.

Group	Baseline	1	2	4	8	16	32
PSNR	31.42	31.62	31.63	31.64	31.77	31.80	31.81
Params	6.81	6.81	6.82	6.82	6.84	6.87	6.94

Table 7: The number of groups in SAM. The results are obtained without using the convolution in Eq. 8.

Method	Gourp conv	w/o Outconv	w/ Outconv
PSNR	31.52	31.80	31.82

Table 8: Different design choices of SAM.

where there are total 8 residual blocks in each scale of both encoder and decoder. To save training time, we train this small model for only 1900 epochs on GoPro (Nah, Hyun Kim, and Mu Lee 2017) to discuss the influence of each design.

Effects of Individual Modules

We study the effects of the proposed two modules individually in Table 3. Compared to the baseline model (Cho et al. 2021), SAM significantly boosts the performance by 0.40 dB. Equipped with FAM, the network obtains 32.17 dB PSNR, which is 0.75 dB higher than that of baseline. These results are achieved with negligible introduced parameters and computational complexity. With both two modules, the model receives a further improved accuracy that is 0.87 dB higher than baseline by introducing only 0.12 M parameters and 0.4 G FLOPs. These results demonstrate the effectiveness of the proposed modules and their compatibility.

Design Choices for SAM

In Table 4, we substitute various alternatives for the activation function used in SAM. With Softmax function that is adopted in the vanilla self-attention, SAM obtains 31.66 dB PSNR. This result is similar to that of Sigmoid function which also projects the attention weights into (0, 1). Interestingly, without using any activation function, the network shows no decline in performance. Equipped with hyperbolic tangent function, SAM receives a remarkable gain of 0.11 dB over Softmax version. In this case, the attention module is capable of suppressing the deleterious information when performing integration within a local region.

In Table 7, we vary the number of groups in SAM and study its influence on accuracy. Increasing the number of groups leads to improved performance consistently. This phenomenon demonstrates that, with various attention weights, the features can be projected into more subspaces corresponding to the case of more heads in MHSA. However, as the number of groups goes up, the number of parameters also increases. Since there is no significant difference in accuracy between group 16 and 32, we select 16 as the final configuration.

To further demonstrate the validity of our spatial attention mechanism, we substitute the standard group convolution for SAM. The main difference between these two methods is that the attention weights of SAM are context-aware and adjust according to the input feature. From Table 8 we can see that using group convolution leads to 0.28 dB PSNR decline compared to our version. Then, we also study the importance of convolution in Eq. 8 which promotes interactions between different groups. As shown in Table 8, this operation boosts the performance by 0.02 dB.

Alternatives to SAM

To demonstrate the effectiveness of our spatial attention module, we choose the vanilla self-attention (Wang et al. 2018) and MDTA in Restormer (Zamir et al. 2022) as competitors. To achieve a fair comparison, we implement them and our SAM in the same network (Cho et al. 2021). Due to the huge memory consumption of the original self-attention, we only use single attention block in the first scale of the decoder network. The results are shown in Table 9. Compared to MDTA in Restormer (Zamir et al. 2022) and vanilla self-attention (Wang et al. 2018) module, our SAM has fewest parameters and FLOPs. Implementing attention on channel dimension, MDTA obtains the worst performance compared to the explicit spatial attention versions. Compared to the global self-attention, SAM receives a gain of 0.05 dB due to the properties of local and selective integration. It is worth mentioning that SAM has less computational complexity

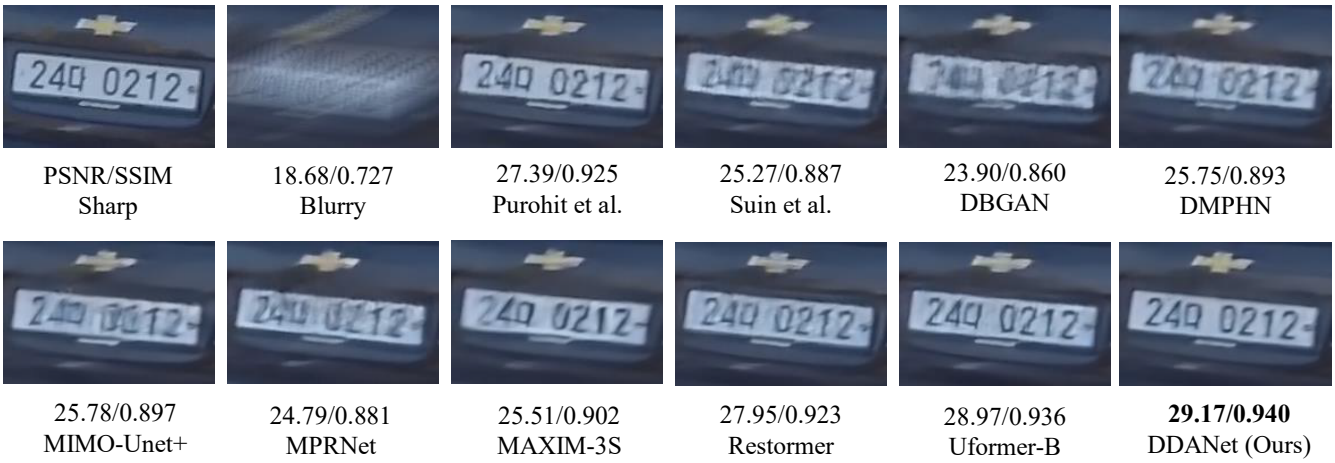

Figure 4: Visual comparisons on GoPro (Nah, Hyun Kim, and Mu Lee 2017) test dataset. Our model is more effective in recovering details than other algorithms.

Method	MDTA	Self-attention	SAM
PSNR (dB)	31.34	31.50	31.55
Params (M)	6.88	8.99	6.83
FLOPs (G)	67.45	67.30	67.17

Table 9: Comparisons between different variants of self-attention. MDTA denotes the multi-Dconv head transposed attention module proposed in Restormer (Zamir et al. 2022). The global self-attention module is borrowed from (Wang et al. 2018).

Filters	Baseline	1	4	6	8
PSNR (dB)	31.42	31.48	31.61	31.66	31.69
FLOPs (G)	67.17	67.18	67.20	67.22	67.23

Table 10: Number of local filters. Four filters indicates that we deploy the local filter to the last four residual blocks in each scale of the baseline network.

compared to self-attention, and thus it can be deployed in multiple locations.

Design Choices for FAM

Since the global branch and local branch of FAM only differ in kernel size, we only explore the influence of the number of local branches in Table 10. The number 4 means that the last four residual blocks of each scale are equipped with this operator. Increasing the number of frequency attention blocks leads to consistent improvement in accuracy. Nevertheless, using more filters means the increasing complexities. To strike a better trade-off between efficiency and accuracy, we pick 4 local filters in the final network. To delve into the mechanism of FAM, we plot the proportion of high-frequency component in the feature map in Figure 5. These results are obtained from the feature before the second convolution of each residual block. With FAM, the high-frequency components are accentuated, according with the aim of image deblurring.

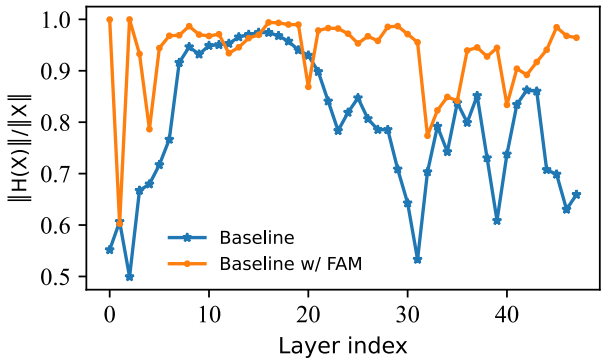

Figure 5: The proportion of high-frequency component in the feature map with different depths. FAM emphasizes the high-frequency parts significantly compared to the baseline network.

Conclusion

To address image deblurring, we develop the dual-domain attention mechanism that is composed of two attention modules in both spatial and frequency domains to enhance representation learning capability. Specifically, the spatial attention module (SAM) mimics the self-attention in an efficient style to reduce computation complexity. SAM generates selective attention weights from the global feature and performs attention in the local region. Furthermore, a simple yet effective frequency attention module (FAM) is proposed to accentuate the useful frequency subbands by decoupling and modulating frequency components . These two modules are incorporated into a U-shaped baseline network. Extensive experiments on the widely used benchmarks demonstrate that DDANet achieves the state-of-the-art accuracy with fast inference speed.

References

Aittala, M.; and Durand, F. 2018. Burst image deblurring using permutation invariant convolutional neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 731–747.

Ayers, G.; and Dainty, J. C. 1988. Iterative blind deconvolution method and its applications. *Optics letters*, 13(7): 547–549.

Bahat, Y.; Efrat, N.; and Irani, M. 2017. Non-uniform blind deblurring by reblurring. In *Proceedings of the IEEE international conference on computer vision*, 3286–3294.

Bertero, M.; Boccacci, P.; and De Mol, C. 2021. *Introduction to inverse problems in imaging*. CRC press.

Campisi, P.; and Egiazarian, K. 2017. *Blind image deconvolution: theory and applications*. CRC press.

Cao, J.; Li, Y.; Sun, M.; Chen, Y.; Lischinski, D.; Cohen-Or, D.; Chen, B.; and Tu, C. 2022. Do-conv: Depthwise over-parameterized convolutional layer. *IEEE Transactions on Image Processing*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021a. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12299–12310.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. *arXiv preprint arXiv:2204.04676*.

Chen, L.; Lu, X.; Zhang, J.; Chu, X.; and Chen, C. 2021b. HINet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 182–192.

Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; and Feng, J. 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3435–3444.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4641–4650.

Cui, Y.; Tao, Y.; Bing, Z.; Ren, W.; Gao, X.; Cao, X.; Huang, K.; and Knoll, A. 2023. Selective Frequency Network for Image Restoration. In *The Eleventh International Conference on Learning Representations*.

Delbracio, M.; and Sapiro, G. 2015. Burst deblurring: Removing camera shake through fourier burst accumulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2385–2393.

Keuper, M.; Schmidt, T.; Temerinac-Ott, M.; Padeken, J.; Heun, P.; Ronneberger, O.; and Brox, T. 2013. Blind deconvolution of widefield fluorescence microscopic data by regularization of the optical transfer function (OTF). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2179–2186.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Kundur, D.; and Hatzinakos, D. 1996. Blind image deconvolution. *IEEE signal processing magazine*, 13(3): 43–64.

Lee, H.; Choi, H.; Sohn, K.; and Min, D. 2022. KNN Local Attention for Image Restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2139–2149.

Li, J.; Xia, X.; Li, W.; Li, H.; Wang, X.; Xiao, X.; Wang, R.; Zheng, M.; and Pan, X. 2022a. Next-ViT: Next Generation Vision Transformer for Efficient Deployment in Realistic Industrial Scenarios. *arXiv preprint arXiv:2207.05501*.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022b. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4804–4814.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Liu, K.-H.; Yeh, C.-H.; Chung, J.-W.; and Chang, C.-Y. 2020. A motion deblur method based on multi-scale high frequency residual image learning. *IEEE Access*, 8: 66025–66036.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lucy, L. B. 1974. An iterative technique for the rectification of observed distributions. *The astronomical journal*, 79: 745.

Mao, X.; Liu, Y.; Shen, W.; Li, Q.; and Wang, Y. 2021. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*.

Michailovich, O. V.; and Adam, D. 2005. A novel approach to the 2-D blind deconvolution problem in medical ultrasound. *IEEE transactions on medical imaging*, 24(1): 86–104.

Nah, S.; Hyun Kim, T.; and Mu Lee, K. 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3883–3891.

Purohit, K.; and Rajagopalan, A. 2020. Region-adaptive dense network for efficient motion deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11882–11889.

Purohit, K.; Suin, M.; Rajagopalan, A.; and Boddeti, V. N. 2021. Spatially-adaptive image restoration using distortion-guided networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2309–2319.

Qin, Z.; Zhang, P.; Wu, F.; and Li, X. 2021. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 783–792.

Richardson, W. H. 1972. Bayesian-based iterative method of image restoration. *JoSA*, 62(1): 55–59.

Rim, J.; Lee, H.; Won, J.; and Cho, S. 2020. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, 184–201. Springer.

Sayed, M.; and Brostow, G. 2021. Improved handling of motion blur in online object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1706–1716.

Shen, Z.; Wang, W.; Lu, X.; Shen, J.; Ling, H.; Xu, T.; and Shao, L. 2019. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5572–5581.

Sroubek, F.; and Flusser, J. 2005. Multichannel blind deconvolution of spatially misaligned images. *IEEE Transactions on Image Processing*, 14(7): 874–883.

Suin, M.; Purohit, K.; and Rajagopalan, A. 2020. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3606–3615.

Tsai, F.-J.; Peng, Y.-T.; Lin, Y.-Y.; Tsai, C.-C.; and Lin, C.-W. 2022. Stripformer: Strip Transformer for Fast Image Deblurring. *arXiv preprint arXiv:2204.04627*.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5769–5780.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, R.; and Tao, D. 2014. Recent progress in image deblurring. *arXiv preprint arXiv:1409.6838*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17683–17693.

Wiener, N.; Wiener, N.; Mathematician, C.; Wiener, N.; Wiener, N.; and Mathématicien, C. 1949. *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*, volume 113. MIT press Cambridge, MA.

Wu, F.; Fan, A.; Baevski, A.; Dauphin, Y. N.; and Auli, M. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Yuan, Y.; Su, W.; and Ma, D. 2020. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3555–3564.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5728–5739.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14821–14831.

Zhang, K.; Ren, W.; Luo, W.; Lai, W.-S.; Stenger, B.; Yang, M.-H.; and Li, H. 2022a. Deep image deblurring: A survey. *International Journal of Computer Vision*, 1–28.

Zhang, Y.; Li, Q.; Qi, M.; Liu, D.; Kong, J.; and Wang, J. 2022b. Multi-scale frequency separation network for image deblurring. *arXiv preprint arXiv:2206.00798*.

Zou, W.; Jiang, M.; Zhang, Y.; Chen, L.; Lu, Z.; and Wu, Y. 2021. Sdwnet: A straight dilated network with wavelet transformation for image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1895–1904.