DFMDA-Net: Dense Fusion and Multi-dimension Aggregation Network for Image Restoration

Huibin Yan , **Shuoyao Wang***
Shenzhen University, Shenzhen, China
huibinyan2020@email.szu.edu.cn, sywang@szu.edu.cn

Abstract

The U-shape (encoder-decoder) architecture, combined with effective blocks, has shown significant success in image restoration. In U-shape models, there is insufficient focus on the feature fusion problem between encoder and decoder features at the same level. Current methods often employ simplistic operations like summation or concatenation, which makes it difficult to strike a balance between performance and complexity. To address this issue, we propose a compression-in-the-middle mechanism, termed Integration-Compression-Integration (ICI), which effectively conducts dense fusion and avoids information loss. From the block design perspective, we design a multi-dimension aggregation (MDA) mechanism, capable of effectively aggregating features from both the channel and spatial dimension. Combining the Integration-Compression-Integration feature fusion and the multi-dimension aggregation, our dense fusion and multi-dimension aggregation network (DFMDA-Net) achieves superior performance over state-of-the-art algorithms on 16 benchmarking datasets for numerous image restoration tasks.

1 Introduction

Bad weather or physical limitations of the camera can degrade the quality of captured images and further negatively impact the robustness of downstream tasks. Image restoration aims to eliminate those annoying degradation (such as noise, rain, and blur), and therefore plays an important role in surveillance, autonomous driving, remote sensing, etc. Due to the ill-posed nature, earlier model-based methods restrict the solution space by relying on handcrafted statistical priors. However, in complex real-world scenarios, it is difficult for these methods to recover faithful results.

With the fast development of deep learning, convolutional neural networks (CNN)- and transformer-based methods have achieved remarkable success by learning generalizable priors from large-scale datasets. In these methods, ingenious architecture or blocks are designed or borrowed from the high-level computation vision tasks. For example, the U-shape (encoder-decoder) architecture [Ronneberger *et al.*, 2015] is widely adopted for the effective hierarchy representation. The blocks are commonly designed based on residual learning and different attention mechanisms. Representative blocks include the residual block [Nah *et al.*, 2017] and transformer block [Zamir *et al.*, 2022].

This paper aims to address the restoration task from both the **U-shape architecture** and the **block design**. In the encoder-decoder architecture, to assist the recovery, the features from the encoder are passed through to the decoder at the same level. For the feature fusion problem, current methods directly conduct the concatenation operation [Wang *et al.*, 2022], the summation operation [Chen *et al.*, 2022a], and first concatenation and then compression [Zamir *et al.*, 2022]. These operations either cause information loss or make the complexity higher[1]. To alleviate this challenge, this paper proposes an Integration-Compression-Integration (ICI) mechanism for feature fusion. Our ICI first conducts integration to avoid information loss, then applies information compression to maintain efficiency, and finally conducts integration again for better information utilization. In this way, our ICI implements effective feature fusion while keeping efficient.

For block design, we revisit existing blocks from the perspective of information aggregation direction. Existing blocks either aggregate features from the spatial dimension such as the residual block [Nah *et al.*, 2017] and window-based attention [Liu *et al.*, 2021] or from the channel dimension such as transposed attention [Zamir *et al.*, 2022], and have achieved remarkable success. Nevertheless, how to effectively aggregate the information from both the channel and spatial dimensions, which can further strengthen the performance, deserves attention[2]. Since the window-based attention shares weights across channels, it is compatible with the transposed attention, as the shared operation limits the diversity of channel features. Moreover, the residual block

*Corresponding author. The authors are also with the Guangdong Province Engineering Laboratory for Digital Creative Technology and the Shenzhen Key Laboratory of Digital Creative Technology.

[1] More details will be illustrated in Section 3.3.

[2] Note that aggregating information from channels in this paper means applying self-attention across channels.

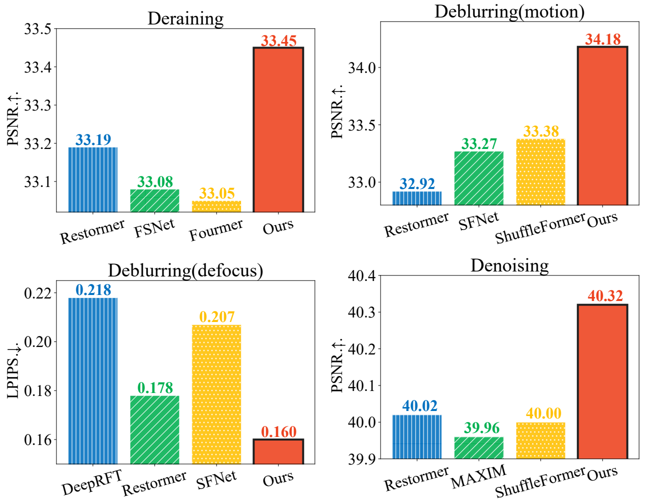

Figure 1: Our method achieves significant performance gain on different image restoration tasks (Tables 2, 3, 4, 7).

and window-based attention can only aggregate the local spatial information, which limits the receptive field. To address the issue, we resort to Fourier frequency learning since the Fourier features embrace global properties and can apply different weights to adjust the frequency points at different channels. To this end, we propose a multi-dimension aggregation (MDA) mechanism, capable of effectively aggregating features from both the channel and spatial dimensions.

By introducing ICI and MDA into a U-shape backbone, we construct the dense fusion and multi-dimension aggregation network (DFMDA-Net). The contributions of this paper can be summarized in four aspects:

- We propose dense fusion and multi-dimension aggregation network for image restoration from the perspective of both the U-shape architecture and block design.
- We propose a dense mechanism for feature fusion, termed Integration-Compression-Integration (ICI). This process is designed to prevent information loss by integration, maintain efficiency through compression in the middle, and thus enhance overall information utilization.
- We propose a multi-dimension aggregation (MDA) mechanism, capable of effectively aggregating features from both the channel and spatial dimension, thus enhancing the restoration performance.
- We conduct extensive experiments on 16 benchmark datasets for numerous image restoration tasks including image deraining, image real denoising, Gaussian image denoising, image motion deblurring, and image defocus deblurring. The results demonstrate that the proposed method achieves state-of-the-art performance, as illustrated in Fig. 1.

2 Related Work

Since image restoration is an ill-posed problem, earlier model-based methods restrict the solution space by exploiting handcrafted statistical priors, such as discriminative prior, channel prior, and gradient prior. These methods are heuristic and usually involve complex optimization problems. Limited by approximate models and the above factors, these methods are usually less effective, especially in complex real-world scenarios.

2.1 CNN-based Methods

With the fast development of deep learning, deep neural networks, especially convolutional neural networks (CNN) have motivated numerous advanced image restoration methods [Cui *et al.*, 2023d; Cui *et al.*, 2023c] by learning dense priors from large-scale datasets. In these methods, there are several excellent convolutional designs which are commonly employed. For example, based on residual learning [He *et al.*, 2016], Resblock [Nah *et al.*, 2017] is proposed and has become the basic module for many image restoration methods. The underlying reason for the effectiveness can be attributed to the fact that the restoration task can be seen as the process of learning the residual between the degraded image and the ground-truth. Spatial and channel attention, which can enhance the desirable features and suppress the detrimental signals, have also achieved remarkable success [Zamir *et al.*, 2021; Cui *et al.*, 2023a]. Moreover, based on U-Net [Ronneberger *et al.*, 2015], the U-shape (encoder-decoder) architecture has become the backbone of most successful image restoration models.

Albeit achieving significant performance gain over model-based methods, CNN-based methods suffer limited receptive field and have poor content adaptation which originates from the basic convolution operation, which is crucial for effective image restoration.

2.2 Transformer-based Methods

Motivated by the great success of transformers [Vaswani *et al.*, 2017], researchers have proposed several excellent restoration models. The self-attention mechanism, which is a core component in transformers [Vaswani *et al.*, 2017], can deal well with the shortcomings of the convolution operation. IPT [Chen *et al.*, 2021a] applies the vanilla transformer on local patches, and has achieved promising performance gain. However, IPT depends on costly pretraining, and cannot be applied to high-resolution images due to the quadratic complexity of self-attention.

To address the quadratic complexity of self-attention, a line of methods applies self-attention on windows [Liu *et al.*, 2021]. For example, SwinIR [Liang *et al.*, 2021] and Uformer [Wang *et al.*, 2022] adopt window-based self-attention (WSA). Although WSA reduces the complexity from quadratic to linear compared to self-attention, it dissevers the global pixel dependencies. Another line of methods applies self-attention across channels instead of spatial dimensions. Notably, Restormer [Zamir *et al.*, 2022] proposes transposed attention (TA) with linear complexity, and has achieved remarkable success in multiple image restoration tasks. In essence, TA is a kind of channel self-attention, which aggregates the channel features with the similarity matrix of channels. Therefore, it still has limited receptive field.

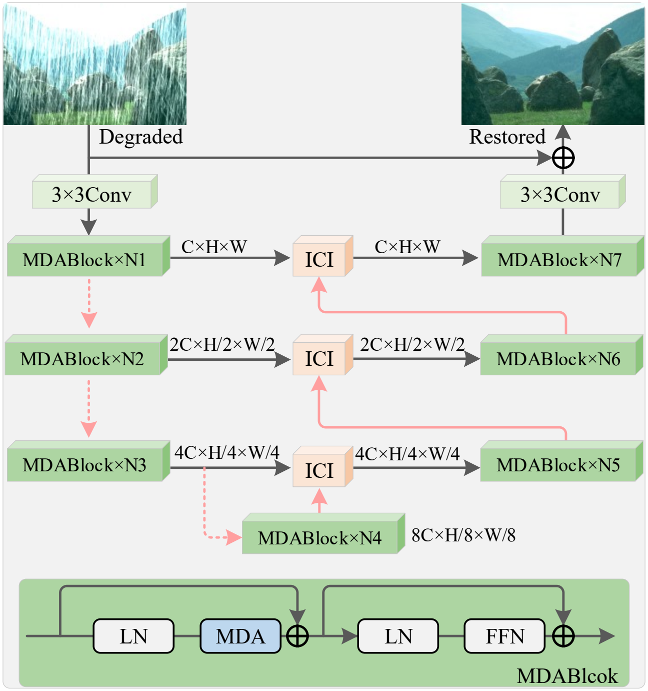

Figure 2: Overall Pipeline. MDABlcok denotes the multi-dimension aggregation block, where MDA is the proposed multi-dimension aggregation mechanism. ICI is the proposed Integration-Compression-Integration mechanism. The red dotted line denotes downsampling. The red solid line represents upsampling.

2.3 Discussions

In this part, we discuss the significance of our work. First, most of the research focuses on exploring effective backbones while ignoring the architectural effects of the U-shape backbone. This paper aims to explore an effective and efficient feature fusion paradigm for the features from the encoder and the decoder at the same level. Second, this paper aims to aggregate desirable features (information) from both the spatial and channel dimensions for effective image restoration. Although Fourier domain learning is not new in image restoration [Mao *et al.*, 2021; Zhou *et al.*, 2023], these methods only achieve the aggregation of spatial information, and fail to adaptively aggregate the features from the channel dimension. The excellent performance of our work demonstrates that effectively aggregating information from both channel and spatial dimensions contributes to image restoration, which is expected to motivate more future research.

3 Methodology

3.1 Overall Pipeline

The over pipeline of the proposed dense fusion and multi-dimension aggregation (DFMDA-Net) is given in Fig. 2. Following most image restoration methods [Wang *et al.*, 2022; Zamir *et al.*, 2022], our DFMDA-Net is constructed based on the hierarchical encoder-decoder architecture. The detailed process is described as follows.

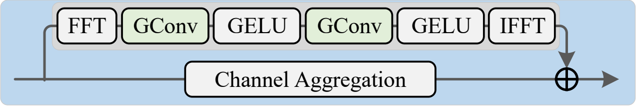

Figure 3: Structure of MDA. FFT and IFFT denote the fast Fourier transform and inverse fast Fourier transform, respectively. GConv denotes the group 1×1 convolution.

Given an input, our DFMDA-Net first applies a 3×3 convolution to obtain the embeddings. Then these embeddings are sequentially passed through a 3-level encoder, a bottleneck, and a 3-level decoder. Each level consists of several multi-dimension aggregation (MDA) blocks. After each level of the encoder, the output features are halved in spatial size and doubled in channel dimension with a 3×3 convolution and the pixel-unshuffle operation. Before each level of the decoder, the input features from the last level are doubled in spatial size and halved in channel dimension with a 3×3 convolution and the pixel-shuffle operation. In addition, to assist the recovery process, the output features from the same level of the encoder and the upsampled features are fused with the proposed Integration-Compression-Integration (ICI) module. Finally, the output features from the last level of the decoder are refined with a 3×3 convolution, and then added to the degraded image to obtain the restored image.

In the following, we introduce the proposed MDA and ICI mechanism.

3.2 Multi-dimension Aggregation

Analysis and Motivation

For block design, we revisit existing blocks from the perspective of information aggregation direction. Existing blocks can be divided into two groups: 1) aggregation from the spatial dimension and 2) aggregation from the channel dimension. CNN-based blocks such as the residual block [Nah *et al.*, 2017], and transformer-based blocks such as recently window-based attention [Liu *et al.*, 2021], lie in category 1. Transposed attention, i.e., applying self-attention across channels, lies in category 2. The effectiveness of these blocks demonstrates that aggregating information from either the spatial or channel direction is vital for effective image restoration. Therefore, joint spatial and channel information aggregation is an intuitive idea to strengthen the performance.

However, directly combining the attention in categories 1 and 2 cannot well accomplish the purpose. Specifically, since the window-based attention shares weights across channels, it is compatible with the transposed attention, as the shared operation limits the diversity of channel features. Moreover, the residual block and window-based attention can only aggregate the local spatial information. Therefore, to effectively integrate information from different dimensions, the spatially aggregated module should be equipped with the global properties and the weights on channels cannot be shared.

To this end, we resort to Fourier frequency learning with two utilizable properties. i) According to the spectral convolution theorem, the Fourier features embrace global properties. Updating a single value in the frequency domain globally affects all original spatial features. ii) Fourier transform

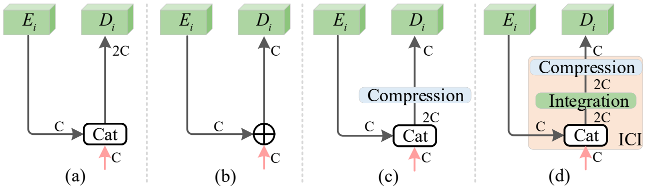

Figure 4: Comparison of feature fusion manner. (a) Integration, (b) Compression-Integration, (c) Compression-Integration, (d) Integration-Compression-Integration (ICI, Ours). E_i represents the i-level encoder. D_i represents the i-level decoder, which consists of several stacked blcoks (*e.g.*, transformer blocks), can be seen as integration. Cat denotes the concatenation operation. $\oplus$ represents the summation operation. C represents the number of channels.

can transfer the features into different frequency components, which can be seen as a frequency disentanglement and rearrangement. This property allows us to selectively enhance the necessary frequencies and suppress the detrimental frequencies in different channels.

Solution

Based on the above analysis, we propose a multi-dimension aggregation (MDA) block, capable of effectively aggregating features from both the channel and spatial dimension, as illustrated in Fig. 3. Specifically, we first transfer the input tensor X into Fourier domain with fast Fourier transform. Then we conduct two group 1×1 convolutions to conduct feature transformation. A GELU activation function is followed with each convolution. Finally, we transfer the obtained features back into the original space with inverse Fourier transform, and add the output features to those of the channel aggregation branch [Zamir *et al.*, 2022] to get the result of MDA. Mathematically, the overall process is defined as:

$$\hat{X} = \mathrm{IFFT}\left(\sigma\left(C_g^2\left(\sigma\left(C_g^1\left(\mathrm{FFT}\left(X\right)\right)\right)\right)\right)\right) + X_A, \quad (1)$$

where C_g^1 and C_g^2 are group 1×1 convolutions. σ denotes the GELU activation function. FFT and IFFT denote fast Fourier transform and inverse fast Fourier transform, respectively. X_A is the result of the channel aggregation branch.

3.3 Integration-Compression-Integration Mechanism

In most of the existing U-shape-based image restoration methods, the features from the encoder are passed through to the decoder at the same level. For the feature fusion problem, current methods directly conduct the concatenation operation [Wang *et al.*, 2022], the summation operation [Chen *et al.*, 2022a], and first concatenation and then compression [Zamir *et al.*, 2022], as illustrated in Fig. 4a-c, respectively. Since the parameters and computational cost of these transformer blocks are proportional to C^2, where C is the number of channels, the concatenation manner (Fig. 4a) is parameter and computation costly. For the summation operation (Fig. 4b), it is ineffective since there exists the misaligned problem between the features of the encoder and the decoder. For the third case (Fig. 4c), it is too rough to conduct the direct compression with a 1×1 convolution to reduce channels, resulting in information loss.

	Fig. 4a	Fig. 4c	Fig. 4d (ICI)
Parameters	$\mathcal{O}(4mC^2)$	$\mathcal{O}(mC^2)$	$\mathcal{O}((m+3m_1)C^2)$
MACs	$\mathcal{O}(4mC^2HW)$	$\mathcal{O}(mC^2HW)$	$\mathcal{O}((m+3m_1)C^2HW)$

Table 1: Parameters and MACs comparisons of different feature fusion methods. Assume there are m blocks used for integration. $m_1 \in [0, m]$.

To alleviate this challenge, this paper revisits the existing feature fusion mechanisms in Fig. 4a-c. By deeming the D_i which consists of stacked blocks as integration, we can abstract them into two categories, *i.e.*, Integration (Fig. 4a), and Compression[3]-Integration (Fig. 4b-c). As demonstrated above, the former is parameter and computation costly, and the latter suffers information loss. This paper proposes an Integration-Compression-Integration (ICI) mechanism for feature fusion, as illustrated in Fig. 4d. Our ICI first conducts integration to avoid information loss, then applies information compression to maintain efficiency, and finally conducts integration again for better information utilization. In this way, our ICI implements effective feature fusion while keeping efficient.

4 Relationship with Existing Feature Fusion Methods

Assume there are m blocks used for integration. The parameters and MACs of each block is $\mathcal{O}(C^2)$ and $\mathcal{O}(C^2HW)$. Fig. 4a means that the m blocks are directly applied to integrate the features with $2C$ channels. Fig. 4c means that the m blocks are used to integrate the features with C channels. We omit the analysis of the manner in Fig. 4b, since Fig. 4b can be seen as a particular case of Fig. 4c. ICI applies $m_1 \in [0, m]$ blocks for the first integration, and $m - m_1$ blocks for the second integration (D_i). In other words, m_1 blocks are used to integrate the features with $2C$ channels, and $m - m_1$ blocks are used to integrate the features with C channels. Therefore, the proposed ICI is a dense feature fusion paradigm compared to existing feature fusion manners shown in Fig. 4a-c. We also list the parameters and MACs of different feature fusion methods in Table 1 for better comparison.

5 Experiments

5.1 Setups

Evaluation Metrics

Following most of image restoration literature [Li *et al.*, 2023; Zamir *et al.*, 2022], we adopt PSNR and SSIM as the evaluation metrics. For defocus deblurring, extra metrics MAE and LPIPS are employed. It is worth noting that the PSNR and SSIM values are only calculated on the Y channel in YCbCr space for image deraining. We introduce the datasets in Appendix A.

Implementation Details

For different image restoration tasks, we train separate models. Unless otherwise specified, all the experiments are with

[3] Summation can be regarded as a kind of compression.

	Test100		Rain100H		Rain100L		Test2800		Test1200		Average	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
PreNet [Ren *et al.*, 2019]	24.81	0.851	26.77	0.858	32.44	0.950	31.75	0.916	31.36	0.911	29.42	0.897
MSPFN [Jiang *et al.*, 2020]	27.50	0.876	28.66	0.860	32.40	0.933	32.82	0.930	32.39	0.916	30.75	0.903
MPRNet [Zamir *et al.*, 2021]	30.27	0.897	30.41	0.890	36.40	0.965	33.64	0.938	32.91	0.916	32.73	0.921
HINet [Chen *et al.*, 2021b]	30.29	0.906	30.65	0.894	37.28	0.970	33.91	0.941	33.05	0.919	33.03	0.926
SPAIR [Purohit *et al.*, 2021]	30.35	0.909	30.95	0.892	36.93	0.969	33.34	0.936	33.04	0.922	32.91	0.926
Restormer [Zamir *et al.*, 2022]	32.00	0.923	31.46	0.904	38.99	**0.978**	34.18	0.944	33.19	0.926	33.96	0.935
MAXIM-2S [Tu *et al.*, 2022]	31.17	0.922	30.81	0.903	38.06	0.977	33.80	0.943	32.37	0.922	33.24	0.933
IRNext [Cui *et al.*, 2023c]	31.53	0.919	31.64	0.902	38.14	0.972	-	-	-	-	-	-
Fourmer [Zhou *et al.*, 2023]	30.54	0.911	30.76	0.896	37.47	0.970	-	-	33.05	0.919	-	-
FSNet [Cui *et al.*, 2023b]	31.05	0.919	31.77	0.906	38.00	0.972	33.64	0.936	33.08	0.916	33.51	0.930
DFMDA-Net (Ours)	**32.32**	**0.928**	**32.01**	**0.911**	**39.15**	0.977	**34.21**	**0.946**	**33.45**	**0.928**	**34.23**	**0.938**

Table 2: **Single image deraining** results.

	GoPro		HIDE		RealBlur-R		RealBlur-J		Average	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SRN [Tao *et al.*, 2018]	30.26	0.934	28.36	0.915	35.66	0.947	28.56	0.867	30.71	0.915
DBGAN [Zhang *et al.*, 2020b]	31.10	0.942	28.94	0.915	33.78	0.909	24.93	0.745	-	-
Suin et al. [Suin *et al.*, 2020]	31.85	0.948	29.98	0.930	-	-	-	-	-	-
SPAIR [Purohit *et al.*, 2021]	32.06	0.953	30.29	0.931	-	-	28.81	0.875	-	-
MIMO-UNet+ [Cho *et al.*, 2021]	32.45	0.957	29.99	0.930	35.54	0.947	27.63	0.837	-	-
IPT [Chen *et al.*, 2021a]	32.52	-	-	-	-	-	-	-	-	-
MPRNet [Zamir *et al.*, 2021]	32.66	0.959	30.96	0.939	35.99	0.952	28.70	0.873	32.08	0.931
Uformer [Wang *et al.*, 2022]	33.06	0.962	30.90	0.940	36.19	0.956	29.09	0.886	32.31	0.936
Restormer [Zamir *et al.*, 2022]	32.92	0.961	31.22	0.942	36.19	0.957	28.96	0.879	32.32	0.935
SFNet [Cui *et al.*, 2023d]	33.27	0.963	31.10	0.941	-	-	-	-	-	-
GRL [Li *et al.*, 2023]	33.93	0.968	31.65	0.947	-	-	-	-	-	-
DiffIR [Xia *et al.*, 2023]	33.20	0.963	31.55	0.947	-	-	-	-	-	-
ShuffleFormer [Xiao *et al.*, 2023]	33.38	0.965	31.25	0.943	**36.34**	**0.958**	**29.19**	**0.890**	32.54	0.939
DFMDA-Net (Ours)	**34.18**	**0.969**	**32.13**	**0.950**	36.31	**0.958**	29.14	0.887	**32.94**	**0.941**

Table 3: **Single image deblurring** results. The model is trained on GoPro, and directly applied to other datasets.

the following settings. The number of blocks [N1, N2, N3, N4, N5, N6, N7] is set to [4, 6, 6, 8, 6, 6, 4]. The number of attention heads and groups of 1×1 convolution in MDA are both set to [1, 2, 4, 8, 4, 2, 1]. The initial channel dimension is 48. The channel expansion ratio in FFN is set to 2.66. m_1 is set to 1. We adopt the progressive learning strategy [Zamir *et al.*, 2022] to train our models for 300K iterations with L1 loss. The optimizer is AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 1×10^{-4}). The initial learning rate is 3×10^{-4}, and declines to 1×10^{-6} with the cosine annealing. For data augmentation, vertical and horizontal flips are employed. The visual results are given in Appendix B.

5.2 Results on Image Deraining

The image deraining results on five datastets (Test100 [Zhang *et al.*, 2020a], Rain100H [Yang *et al.*, 2017], Rain100L [Yang *et al.*, 2017], Test2800 [Fu *et al.*, 2017], and Test1200 [Zhang and Patel, 2018]) are given in Table 2. On all five datasets, our DFMDA-Net achieves the best results on PSNR. Compared with the best transformer-based method, Restormer, our DFMDA-Net brings 0.27 dB PSNR improvement on average. When paying attention to the individual dataset Rain100H, our DFMDA-Net achieves 0.55 dB PSNR improvement. Compared with the frequency method Fourmer, our DFMDA-Net achieves a performance gain of more than 1 dB on three datasets including Test100, Rain100H, and Rain100L. Moreover, compared with the recent method FSNet, our DFMDA-Net brings 0.72 dB PSNR improvement on average. When paying to the individual datasets Test100 and Rain100L, the improvements are up to 1.27 dB and 1.15 dB, respectively.

5.3 Results on Motion Deblurring

We gives the image motion deblurring results on datastets (GoPro [Nah *et al.*, 2017], HIDE [Shen *et al.*, 2019], and Realbur [Rim *et al.*, 2020] (Realbur-R and Realbur-J)) in Table 3. Compared with Restormer, our DFMDA-Net brings the PSNR improvement as large as 0.62 dB on average. When averaged across all the datasets, our DFMDA-Net achieves 0.40 dB PSNR improvement over the global ShuffleFormer. Moreover, compared with the global modeling method, GRL, which simultaneously models the pixel, region, and global level interactions, our DFMDA-Net achieves 0.25 dB and 0.48 dB improvements on GoPro and HIDE, respectively.

5.4 Results on Defocus Deblurring

Table 4 presents the results of image defocus deblurring on the DPDD [Abuolaim and Brown, 2020] dataset. Our

	Indoor Scenes				Outdoor Scenes				Combined			
Method	PSNR	SSIM	MAE	LPIPS	PSNR	SSIM	MAE	LPIPS	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓
DMENet [Lee *et al.*, 2019]	25.50	0.788	0.038	0.298	21.43	0.644	0.063	0.397	23.41	0.714	0.051	0.349
DPDNet [Abuolaim and Brown, 2020]	26.54	0.816	0.031	0.239	22.25	0.682	0.056	0.313	24.34	0.747	0.044	0.277
KPAC [Son *et al.*, 2021]	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227
IFAN [Lee *et al.*, 2021]	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217
DeepRFT [Mao *et al.*, 2021]	-	-	-	-	-	-	-	-	25.71	0.801	0.039	0.218
DRBNet [Ruan *et al.*, 2022]	-	-	-	-	-	-	-	-	25.73	0.791	-	0.183
Restormer [Zamir *et al.*, 2022]	28.87	0.882	0.025	0.145	23.24	0.743	0.050	0.209	25.98	0.811	0.038	0.178
NRKNet [Quan *et al.*, 2023]	-	-	-	-	-	-	-	-	26.11	0.810	-	0.210
GRL [Li *et al.*, 2023]	29.06	0.886	0.024	0.139	23.45	**0.761**	0.049	0.196	26.18	0.822	0.037	0.168
FocalNet [Cui *et al.*, 2023a]	29.10	0.876	0.024	0.173	23.41	0.743	0.049	0.246	26.18	0.808	0.037	0.210
FSNet [Cui *et al.*, 2023b]	29.14	0.878	0.024	0.166	23.45	0.747	0.050	0.246	26.22	0.811	0.037	0.207
DFMDA-Net (Ours)	**29.35**	**0.891**	**0.024**	**0.129**	**23.52**	0.760	**0.049**	**0.189**	**26.36**	**0.824**	**0.037**	**0.160**

Table 4: **Single image defocus deblurring** results on the DPDD dataset.

	Set12			BSD68			Urban100		
Method	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
DnCNN [Zhang *et al.*, 2017a]	32.67	30.35	27.18	31.62	29.16	26.23	32.28	29.80	26.35
FFDNet [Zhang *et al.*, 2018]	32.75	30.43	27.32	31.63	29.19	26.29	32.40	29.90	26.50
IRCNN [Zhang *et al.*, 2017b]	32.76	30.37	27.12	31.63	29.15	26.19	32.46	29.80	26.22
DRUNet [Zhang *et al.*, 2021]	33.25	30.94	27.90	31.91	29.48	26.59	33.44	31.11	27.96
Restormer [Zamir *et al.*, 2022]	33.35	31.04	28.01	31.95	29.51	26.62	33.67	31.39	28.33
DFMDA-Net (Ours)	**33.39**	**31.09**	**28.05**	**31.96**	**29.52**	**26.63**	**33.82**	**31.60**	**28.62**

Table 5: **Gaussian grayscale image denoising** results.

	CBSD68			Kodak24			McMaster			Urban100		
Method	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50	σ=15	σ=25	σ=50
IRCNN [Zhang *et al.*, 2017b]	33.86	31.16	27.86	34.69	32.18	28.93	34.58	32.18	28.91	33.78	31.20	27.70
FFDNet [Zhang *et al.*, 2018]	33.87	31.21	27.96	34.63	32.13	28.98	34.66	32.35	29.18	33.83	31.40	28.05
DnCNN [Zhang *et al.*, 2017a]	33.90	31.24	27.95	34.60	32.14	28.95	33.45	31.52	28.62	32.98	30.81	27.59
DSNet [Peng *et al.*, 2019]	33.91	31.28	28.05	34.63	32.16	29.05	34.67	32.40	29.28	-	-	-
DRUNet [Zhang *et al.*, 2021]	34.30	31.69	28.51	35.31	32.89	29.86	35.40	33.14	30.08	34.81	32.60	29.61
Restormer [Zamir *et al.*, 2022]	34.39	31.78	28.59	35.44	33.02	30.00	35.55	33.31	30.29	35.06	32.91	30.02
DFMDA-Net (Ours)	**34.40**	**31.79**	**28.60**	**35.49**	**33.07**	**30.06**	**35.62**	**33.39**	**30.35**	**35.22**	**33.11**	**30.29**

Table 6: **Gaussian color image denoising** results.

DFMDA-Net achieves basically consistent performance gain on both scene categories. Compared with Restormer, our DFMDA-Net brings 0.38 dB PSNR improvement on the combined category. When paying attention to the indoor scene category, our DFMDA-Net achieves 0.48 dB PSNR improvement. Compared with GRL, our DFMDA-Net achieves a performance gain of 0.18 dB on the combined category. Moreover, compared with the recent FSNet, our DFMDA-Net brings 0.047 LPIPS improvement on average. When paying to the outdoor scene category, the improvement is up to 0.057.

5.5 Results on Image Denoising

For image denoising, we conduct the experiments on both synthetic and real datasets.

Gaussian Image Denoising

For Gaussian image denoising, we generate the synthetic benchmark datasets by adding additive white Gaussian noise on DIV2K [Agustsson and Timofte, 2017], Flickr2K, BSD500 [Arbelaez *et al.*, 2010] and WED [Ma *et al.*, 2016]. The testing datasets are Set12 [Zhang *et al.*, 2017a], BSD68 [Martin *et al.*, 2001], Urban100 [Huang *et al.*, 2015], Kodak24, and McMaster [Zhang *et al.*, 2011]. We list the results in Table 5 for grayscale images and Table 6 for color images. Following DRUNet [Zhang *et al.*, 2021], three noise levels (15, 25, and 50) are tested. Overall, our DFMDA-Net has consistent performance gain on different datasets and noise levels. Compared with Restormer, our DFMDA-Net brings 0.29 dB and 0.27 dB PSNR improvement on the challenging 50 noise level on grayscale and color images of Urban100, respectively.

Real Image Denoising

We conduct the real image denoising on the SIDD [Abdelhamed *et al.*, 2018] dataset, and the result is shown in Table 7. Compared with the recent best method, Restormer, our DFMDA-Net achieves 0.30 dB PSNR improvement. Compared with the recent global model, ShuffleFormer, our DFMDA-Net brings 0.32 dB PSNR gain.

Method	DnCNN [Zhang *et al.*, 2017a]	BM3D [Dabov *et al.*, 2007]	VDN [Yue *et al.*, 2019]	MIRNet [Zamir *et al.*, 2020]	MPRNet [Zamir *et al.*, 2021]	HINet [Chen *et al.*, 2021b]	NBNet [Cheng *et al.*, 2021]
PSNR	23.66	25.65	39.28	39.72	39.71	39.99	39.75
SSIM	0.583	0.685	0.956	0.959	0.958	0.958	0.959
Method	DAGL [Mou *et al.*, 2021]	Uformer [Wang *et al.*, 2022]	Restormer [Zamir *et al.*, 2022]	MAXIM-3S [Tu *et al.*, 2022]	CAT [Chen *et al.*, 2022b]	ShuffleFormer [Xiao *et al.*, 2023]	DFMDA-Net Ours
PSNR	38.94	39.89	40.02	39.96	40.00	40.01	**40.32**
SSIM	0.953	0.960	0.960	0.960	0.960	0.960	**0.963**

Table 7: **Real image denoising** results on the SIDD dataset.

Method	MDA	ICI	PSNR	Params (M)	MACs (G)
baseline	×	×	33.16	11.74	64.29
baseline1	✓	×	33.34	12.10	70.54
baseline2	×	✓	33.26	12.49	69.23
Ours	✓	✓	33.45	12.86	75.58

Table 8: Ablation study for each module.

m_1	0	1	2	3	4	5	6
PSNR	33.34	33.45	33.47	33.48	33.49	33.49	33.50
Params (M)	12.10	12.86	13.62	14.39	15.15	15.91	16.68
MACs (G)	70.54	75.58	80.62	85.66	90.70	95.74	100.78

Table 9: Ablation study for ICI.

Group	1	H	2H	4H
PSNR	33.42	33.45	33.41	33.25
Params (M)	14.23	12.86	12.68	12.59
MACs (G)	79.36	75.58	72.40	70.82

Table 10: Ablation study for MDA.

Method	Fig. 4a	Fig. 4b	Fig. 4c	ICI (Ours)
PSNR	33.50	33.27	33.34	33.45
Params (M)	16.68	12.06	12.10	12.86
MACs (G)	100.78	70.27	70.54	75.58

Table 11: Results of alternatives to ICI.

5.6 Ablation Study

In this section, we first ablate the proposed two modules, ICI and MDA. Then we investigate the design choices of these modules. The models are trained on the GoPro [Nah *et al.*, 2017] dataset for 150K iterations. The initial channel dimension is set to 32. Unless specified otherwise, other model configurations and training settings are the same as those in Section 5.1.

Effectiveness of Each Module

The results are given in Table 8. The baseline model is Restormer [Zamir *et al.*, 2022], and achieves 33.16 dB PSNR. When equipped with MDA and ICI, the models bring 0.18 dB and 0.10 dB performance gains over the baseline model. With the introduction of both MDA and ICI, the model achieves 33.45 dB PSNR. The above results demonstrate the effectiveness of the proposed two modules.

Number of Blocks (m_1) in ICI

We investigate the effects of the number of blocks (m_1) used for the first integration in ICI. Table 9 lists the results. As m_1 increases, the performance boosts rapidly when m_1=1, and then tends to saturate when m_1=4. This phenomenon demonstrates that without the first integration, the performance drops significantly. However, applying all the blocks (m_1=6) is parameter- and complexity-costly. These results demonstrate the effectiveness and necessity of the proposed ICI. Considering the performance and efficiency, we empirically choose m_1=1.

Number of Groups in 1×1 Convolution in MDA

We investigate the effects of the number of groups in 1×1 convolution. The results are given in Table 10. 1 represents that the regular 1×1 convolution is employed. H denotes that the number of groups is the same as the number of heads in self-attention. 2H denotes that the number of groups is twice the number of heads in self-attention, and the same meaning for 4H. As can be seen, the H case achieves the best result.

Alternatives to ICI

We also give the results of different alternatives of ICI in Table 11. As can be seen, our ICI achieves a better trade-off between performance and efficiency.

6 Conclusions

This paper presents a dense fusion and multi-dimension aggregation network (DFMDA-Net) for image restoration. For the fusion of features between the encoder and the decoder at the same level, we propose an Integration-Compression-Integration (ICI) mechanism. ICI effectively conducts dense fusion, and achieves a better trade-off between effectiveness and efficiency. In addition, we also propose a multi-dimension aggregation (MDA) mechanism, capable of effectively aggregating features from both the channel and spatial dimensions. Extensive experiments on 16 benchmark datasets demonstrate that DFMDA-Net achieves state-of-the-art performance on image restoration tasks including image deraining, image motion deblurring, image defocus deblurring, real image denoising, and Gaussian image denoising.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62101336 and in part by the Guangdong Basic and Applied Basic Research Foundation under Grants 2022A1515011301.

References

[Abdelhamed *et al.*, 2018] Abdelrahman Abdelhamed, Stephen Lin, et al. A high-quality denoising dataset for smartphone cameras. In *Proc. of CVPR*, 2018.

[Abuolaim and Brown, 2020] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *Proc. of ECCV*, 2020.

[Agustsson and Timofte, 2017] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proc. of CVPR*, 2017.

[Arbelaez *et al.*, 2010] Pablo Arbelaez, Michael Maire, et al. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 2010.

[Chen *et al.*, 2021a] Hanting Chen, Yunhe Wang, et al. Pre-trained image processing transformer. In *Proc. of CVPR*, 2021.

[Chen *et al.*, 2021b] Liangyu Chen, Xin Lu, et al. Hinet: Half instance normalization network for image restoration. In *Proc. of CVPR*, 2021.

[Chen *et al.*, 2022a] Liangyu Chen, Xiaojie Chu, et al. Simple baselines for image restoration. In *Proc. of ECCV*, 2022.

[Chen *et al.*, 2022b] Zheng Chen, Yulun Zhang, et al. Cross aggregation transformer for image restoration. *Proc. of NeurIPS*, 2022.

[Cheng *et al.*, 2021] Shen Cheng, Yuzhi Wang, et al. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proc. of CVPR*, 2021.

[Cho *et al.*, 2021] Sung-Jin Cho, Seo-Won Ji, et al. Rethinking coarse-to-fine approach in single image deblurring. In *Proc. of ICCV*, 2021.

[Cui *et al.*, 2023a] Yuning Cui, Wenqi Ren, et al. Focal network for image restoration. In *Proc. of ICCV*, 2023.

[Cui *et al.*, 2023b] Yuning Cui, Wenqi Ren, et al. Image restoration via frequency selection. *IEEE TPAMI*, 2023.

[Cui *et al.*, 2023c] Yuning Cui, Wenqi Ren, et al. Irnext: Rethinking convolutional network design for image restoration. In *Proc. of ICML*, 2023.

[Cui *et al.*, 2023d] Yuning Cui, Yi Tao, et al. Selective frequency network for image restoration. In *Proc. of ICLR*, 2023.

[Dabov *et al.*, 2007] Kostadin Dabov, Alessandro Foi, et al. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE TIP*, 2007.

[Fu *et al.*, 2017] Xueyang Fu, Jiabin Huang, et al. Removing rain from single images via a deep detail network. In *Proc. of CVPR*, 2017.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, et al. Deep residual learning for image recognition. In *Proc. of CVPR*, 2016.

[Huang *et al.*, 2015] Jia-Bin Huang, Abhishek Singh, et al. Single image super-resolution from transformed self-exemplars. In *Proc. of CVPR*, 2015.

[Jiang *et al.*, 2020] Kui Jiang, Zhongyuan Wang, et al. Multi-scale progressive fusion network for single image deraining. In *Proc. of CVPR*, 2020.

[Lee *et al.*, 2019] Junyong Lee, Sungkil Lee, et al. Deep defocus map estimation using domain adaptation. In *Proc. of CVPR*, 2019.

[Lee *et al.*, 2021] Junyong Lee, Hyeongseok Son, et al. Iterative filter adaptive network for single image defocus deblurring. In *Proc. of CVPR*, 2021.

[Li *et al.*, 2023] Yawei Li, Yuchen Fan, et al. Efficient and explicit modelling of image hierarchies for image restoration. *Proc. of CVPR*, 2023.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, et al. Swinir: Image restoration using swin transformer. In *Proc. of ICCV*, 2021.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of ICCV*, 2021.

[Ma *et al.*, 2016] Kede Ma, Zhengfang Duanmu, et al. Waterloo exploration database: New challenges for image quality assessment models. *IEEE TIP*, 2016.

[Mao *et al.*, 2021] Xintian Mao, Yiming Liu, et al. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021.

[Martin *et al.*, 2001] David Martin, Charless Fowlkes, et al. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. of ICCV*, 2001.

[Mou *et al.*, 2021] Chong Mou, Jian Zhang, et al. Dynamic attentive graph learning for image restoration. In *Proc. of ICCV*, 2021.

[Nah *et al.*, 2017] Seungjun Nah, Tae Hyun Kim, et al. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proc. of CVPR*, 2017.

[Peng *et al.*, 2019] Yali Peng, Lu Zhang, et al. Dilated residual networks with symmetric skip connection for image denoising. *Neurocomputing*, 2019.

[Purohit *et al.*, 2021] Kuldeep Purohit, Maitreya Suin, et al. Spatially-adaptive image restoration using distortion-guided networks. In *Proc. of ICCV*, 2021.

[Quan *et al.*, 2023] Yuhui Quan, Zicong Wu, et al. Neumann network with recursive kernels for single image defocus deblurring. In *Proc. of CVPR*, 2023.

[Ren *et al.*, 2019] Dongwei Ren, Wangmeng Zuo, et al. Progressive image deraining networks: A better and simpler baseline. In *Proc. of CVPR*, 2019.

[Rim *et al.*, 2020] Jaesung Rim, Haeyun Lee, et al. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *Proc. of ECCV*, 2020.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, et al. U-net: Convolutional networks for biomedical image segmentation. In *Proc. of MICCAI*, 2015.

[Ruan *et al.*, 2022] Lingyan Ruan, Bin Chen, et al. Learning to deblur using light field generated and real defocus images. In *Proc. of CVPR*, 2022.

[Shen *et al.*, 2019] Ziyi Shen, Wenguan Wang, et al. Human-aware motion deblurring. In *Proc. of ICCV*, 2019.

[Son *et al.*, 2021] Hyeongseok Son, Junyong Lee, et al. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proc. of ICCV*, 2021.

[Suin *et al.*, 2020] Maitreya Suin, Kuldeep Purohit, et al. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proc. of CVPR*, 2020.

[Tao *et al.*, 2018] Xin Tao, Hongyun Gao, et al. Scale-recurrent network for deep image deblurring. In *Proc. of CVPR*, 2018.

[Tu *et al.*, 2022] Zhengzhong Tu, Hossein Talebi, et al. Maxim: Multi-axis mlp for image processing. In *Proc. of CVPR*, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, et al. Attention is all you need. In *Proc. of NeurIPS*, 2017.

[Wang *et al.*, 2022] Zhendong Wang, Xiaodong Cun, et al. Uformer: A general u-shaped transformer for image restoration. In *Proc. of CVPR*, 2022.

[Xia *et al.*, 2023] Bin Xia, Yulun Zhang, et al. Diffir: Efficient diffusion model for image restoration. In *Proc. of ICCV*, 2023.

[Xiao *et al.*, 2023] Jie Xiao, Xueyang Fu, et al. Random shuffle transformer for image restoration. In *Proc. of ICML*, 2023.

[Yang *et al.*, 2017] Wenhan Yang, Robby T. Tan, et al. Deep joint rain detection and removal from a single image. In *Proc. of CVPR*, 2017.

[Yue *et al.*, 2019] Zongsheng Yue, Hongwei Yong, et al. Variational denoising network: Toward blind noise modeling and removal. *Proc. of NeurIPS*, 2019.

[Zamir *et al.*, 2020] Syed Waqas Zamir, Aditya Arora, et al. Learning enriched features for real image restoration and enhancement. In *Proc. of ECCV*, 2020.

[Zamir *et al.*, 2021] Syed Waqas Zamir, Aditya Arora, et al. Multi-stage progressive image restoration. In *Proc. of CVPR*, 2021.

[Zamir *et al.*, 2022] Syed Waqas Zamir, Aditya Arora, et al. Restormer: Efficient transformer for high-resolution image restoration. In *Proc. of CVPR*, 2022.

[Zhang and Patel, 2018] He Zhang and Vishal M. Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proc. of CVPR*, 2018.

[Zhang *et al.*, 2011] Lei Zhang, Xiaolin Wu, et al. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *JEI*, 2011.

[Zhang *et al.*, 2017a] Kai Zhang, Wangmeng Zuo, et al. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE TIP*, 2017.

[Zhang *et al.*, 2017b] Kai Zhang, Wangmeng Zuo, et al. Learning deep cnn denoiser prior for image restoration. In *Proc. of CVPR*, 2017.

[Zhang *et al.*, 2018] Kai Zhang, Wangmeng Zuo, et al. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE TIP*, 2018.

[Zhang *et al.*, 2020a] He Zhang, Vishwanath Sindagi, et al. Image de-raining using a conditional generative adversarial network. *IEEE TCSVT*, 2020.

[Zhang *et al.*, 2020b] Kaihao Zhang, Wenhan Luo, et al. Deblurring by realistic blurring. In *Proc. of CVPR*, 2020.

[Zhang *et al.*, 2021] Kai Zhang, Yawei Li, et al. Plug-and-play image restoration with deep denoiser prior. *IEEE TPAMI*, 2021.

[Zhou *et al.*, 2023] Man Zhou, Jie Huang, et al. Fourmer: an efficient global modeling paradigm for image restoration. In *Proc. of ICML*, 2023.