Revitalizing Convolutional Network for Image Restoration

Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll, *Fellow, IEEE*

Abstract—Image restoration aims to reconstruct a high-quality image from its corrupted version, playing essential roles in many scenarios. Recent years have witnessed a paradigm shift in image restoration from convolutional neural networks (CNNs) to Transformer-based models due to their powerful ability to model long-range pixel interactions. In this paper, we explore the potential of CNNs for image restoration and show that the proposed simple convolutional network architecture, termed ConvIR, can perform on par with or better than the Transformer counterparts. By re-examing the characteristics of advanced image restoration algorithms, we discover several key factors leading to the performance improvement of restoration models. This motivates us to develop a novel network for image restoration based on cheap convolution operators. Comprehensive experiments demonstrate that our ConvIR delivers state-of-the-art performance with low computation complexity among 20 benchmark datasets on five representative image restoration tasks, including image dehazing, image motion/defocus deblurring, image deraining, and image desnowing.

Index Terms—Convolutional neural networks, frequency modulation, image restoration, representation learning

1 Introduction

As one of the most fundamental vision tasks, image restoration aims to restore a clean image from its degraded counterpart, playing an important role in remote sensing, unmanned systems, photography, and medical imaging [1]–[3]. Due to the ill-posedness of this inverse problem, many conventional algorithms have been developed based on hand-crafted features to reduce the solution space, which are impractical for real-world scenarios [4].

With the development of deep learning, manifold CNN-based frameworks have been proposed based on ingenious modules or borrowed units, such as encoder-decoder architecture [5], [6], dilated convolution [7]–[9], dense connection [10], and attention mechanisms [11], [12]. Recent years have witnessed a paradigm shift from CNN-based architectures to Transformer models [13], [14]. These models have significantly advanced state-of-the-art performance of image restoration by providing long-range pixel interactions and adaptivity ability regarding input features. Despite a few remedies [15]–[17], reducing the complexity of self-attention for image restoration is still a non-trivial problem.

The main goal of this paper is to exploit an efficient and effective image restoration architecture based on CNNs, which can perform on par with or better than Transformer models. By delving into previous advanced image restoration approaches, we summarize several critical factors that a successful image restoration model has: **(a)** Multi-scale representation learning. Recent deep architectures resort to a single encoder-decoder [6], [18], [19] or multi-stage paradigm [9], [12], [20] to learn multi-scale feature representations, which help remove degradation blurs of different sizes. **(b)** Spatial attention. Spatial attention facilitates models to attend to the important region, which is useful for handling spatially-varying blurs [11], [21], [22]. **(c)** Frequency modulation. Frequency modulation operation is a powerful complement to the spatial feature refinement by reducing the frequency discrepancy between sharp and degraded image pairs [23], [24]. **(d)** Low computational complexity. This is essential for image restoration, which often involves high-resolution images.

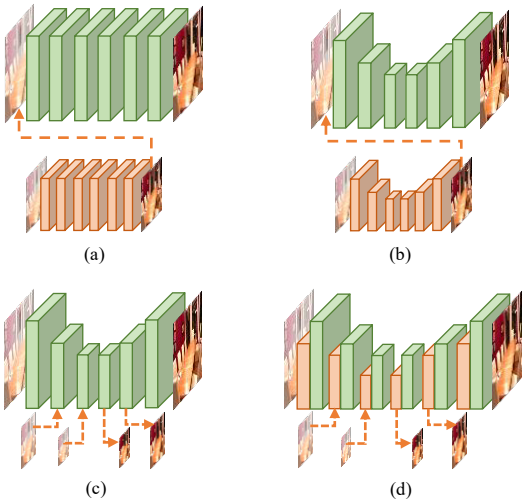


Fig. 1. Comparisons between multi-scale architectures. **(a)** Multi-stage isotropic design [9], [25]. **(b)** Multi-stage U-shaped design [26]. **(c)** Single U-shaped design [6], [18]. **(d)** Our network that imitates the multi-stage design in a U-shaped pipeline.

Considering the above analyses, we rethink the design of convolutional networks and develop an efficient and effective architecture for image restoration. **Firstly**, towards multi-scale learning, we review several representative multi-scale architectures in Figure 1 and propose

- *Y. Cui and A. Knoll are with School of Computation, Information and Technology, Technical University of Munich, 85748 Garching, Munich, Germany. Email: {yuning.cui, knoll}@in.tum.de.*
- *W. Ren and X. Cao are with School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, No. 66, Gongchang Road, Guangming District, Shenzhen, Guangdong 518107, P.R. China. Email: {renwq3, caoxiaochun}@mail.sysu.edu.cn.*

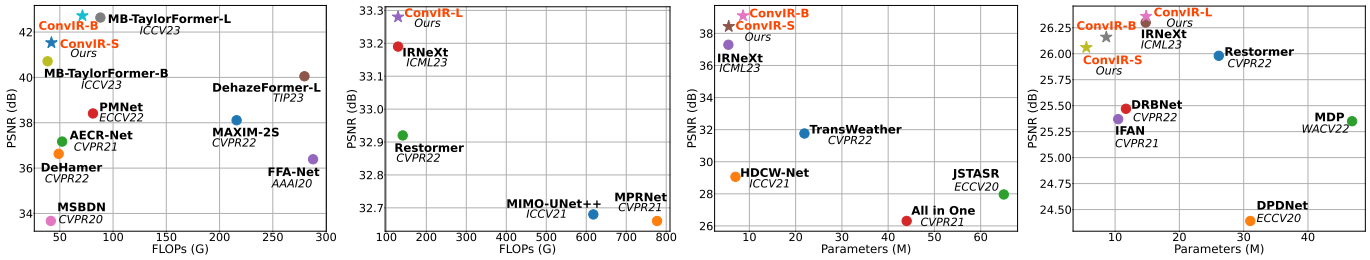


Fig. 2. Comparisons between the proposed models and state-of-the-art algorithms for four image restoration tasks. (a) PSNR *vs.* FLOPs on the SOTS-Indoor [27] dataset for image dehazing; (b) PSNR *vs.* FLOPs on the GoPro [25] dataset for image motion deblurring; (c) PSNR *vs.* parameters on the CSD [28] dataset for image desnowing; and (d) PSNR *vs.* parameters on the DPDD [29] dataset for image defocus deblurring.

imitating the multi-stage mechanism in a single U-shaped network. Specifically, for each scale of the U-shaped architecture, we downsample the feature map into different sizes so that the model can handle blurs in a coarse-to-fine manner. **Secondly**, regarding the attention mechanism design, we develop a multi-shape attention module, which not only can perform efficient information aggregation in multiple shapes, *i.e.*, square and two-directional rectangle regions, but is adaptive to the input feature. Furthermore, this module adopts the idea of dilated convolution to enlarge perception fields. **Thirdly**, due to significant frequency discrepancies between shape and degraded image pairs and the varying importance of different spectral components [30], we accentuate the informative frequency components for high-quality restoration by recalibrating the weight of the high-pass filter in the obtained attention map of the multi-shape attention module. **Finally**, we insert the above modules into a convolutional U-shaped backbone to establish our ConvIR, which obtains comparable or better performance than Transformer models.

For image dehazing, our model outperforms the previous state-of-the-art method (MB-TaylorFormer-L [31]) by 0.08 dB and 1.33 dB PSNR on the SOTS-Indoor [27] and SOTS-Outdoor [27] benchmarks, respectively, with higher efficiency, as illustrated in Figure 2 (a). For the motion blur removal task, our method achieves a performance gain of 0.22 dB PSNR over the recent Transformer-based PromptRestorer [32] on the GoPro [25] dataset. Furthermore, our model shows potential on the desnowing task and outperforms TransWeather [33] by 7.34 dB PSNR on the CSD [28] dataset. Also, on the DPDD [29] dataset for defocus deblurring, our network obtains 26.36 dB PSNR in the combined category, an improvement of 0.38 dB PSNR over Restormer [16].

Overall, the main contributions of this paper can be summarized as follows:

- We identify the properties that a successful image restoration method possesses and propose a novel convolutional model, ConvIR, which enhances multi-scale representation learning by incorporating the multi-stage mechanism into a U-shaped network.
- We present an efficient content-aware dilated multi-shape attention module that can emphasize the informative frequency components by reweighing the weight of the high-pass filter.
- Extensive experiments demonstrate that our model delivers state-of-the-art performance on 20 datasets for five typical image restoration tasks, *i.e.*, image dehazing, image motion/defocus deblurring, image deraining, and image desnowing.

This study is an extension of our conference paper [34]. Compared to the preliminary version, the main improvements of this paper are:

a) **Architectural improvements.** In comparison to the single-shape attention mechanism in the conference paper [34], we develop a multi-shape attention (MSA) module by additionally incorporating a two-directional rectangle attention unit. Moreover, we inject the dilated operation with different rates into the proposed MSA to enlarge the receptive fields. The architectural modifications boost the performance of the model while requiring negligible computation overhead. For example, our model produces a performance gain of 0.09 dB PSNR over IRNeXt [34] on the GoPro [25] dataset with extra only 0.07M parameters and 0.01G FLOPs.

b) **Experiments.** Our model is extended to the nighttime (NHR [35], GTA5 [36]) and remote sensing (SateHaze1k-Thin/Moderate/Thick [37]) image dehazing problems and achieves state-of-the-art performance. Furthermore, we evaluate our model on more synthetic and real-world datasets, such as Haze4K [38], O-HAZE [39], and I-Haze [40]. In addition, we carry out more ablation studies for the key components of our method.

c) **Model diversity**. To promote deployment convenience and demonstrate the effectiveness of our design, we provide three versions of our model, *i.e.*, ConvIR-S (Small), ConvIR-B (Base), and ConvIR-L (Large), for comprehensive comparisons with state-of-the-art algorithms on different problems. In the conference version [34], we mostly adopt ConvIR-L for image deblurring tasks and ConvIR-S for image dehazing and desnowing. It is worth mentioning that our small version still outperforms the strong Transformer-based Restormer [16] with fewer parameters on the DPDD [29] dataset for image defocus deblurring, as illustrated in Figure 2 (d). Also, ConvIR-B achieves state-of-the-art performance on the SOTS-Indoor [27] dataset, producing a performance gain of 0.08 dB PSNR with 19% lower complexity compared to the recent Transformer-based MB-TaylorFormer-L [31], as shown in Figure 2 (a).

2 Related Work

2.1 Image Restoration

As a long-standing task, image restoration provides high-quality images for visibility and downstream high-level tasks, such as object detection [41] and segmentation [42]. CNNs have become the mainstream in this field for several years and achieved many successful stories on various restoration tasks [25], [29], [43], [44]. To boost performance, numerous advanced modules have been developed and borrowed from other domains to strengthen the ability of these CNN-based frameworks [45]. For example, the encoder-decoder architecture is popular for efficient hierarchical representation learning [5], [6], [46]. Multi-stage paradigms [26], [47] and multi-patch learning methods [12], [21], [48] are used to restore clean images in a coarse-to-fine manner. Dilated convolutions help extract multi-scale features and capture large receptive fields [7], [23].

More recently, Transformer models have been introduced in image restoration and have significantly advanced the state-of-the-art performance of various image restoration tasks due to the powerful ability of self-attention to model long-range dependencies [13]. For instance, Guo *et al.* [49] first introduce Transformer into image dehazing. Chen *et al.* [50] propose a multi-scale projection transformer for snow removal. However, the key component of Transformer models, self-attention, has quadratic complexity with respect to the input size. A few remedies alleviate this issue by reducing the attention operation size [51] or switching the operation dimension. Liang *et al.* [52] and Tsai *et al.* [17] compute self-attention within local windows and strip regions, respectively. Zamir *et al.* [16] seek solutions by applying self-attention across channels instead of the spatial dimension. Nonetheless, how to reduce the complexity of Transformer remains an intractable problem in this domain for practical applications.

2.2 Attention Mechanisms

Driven by the success of attention mechanisms in high-level tasks, such as classification and detection, various attention modules have been proposed to attend to essential contents for image restoration [20], [21], [53]. For example, Qin *et al.* [11] combines channel attention and pixel attention mechanisms for image dehazing to treat different features and pixels unequally. Zamir *et al.* [12] devise a supervised attention module to control the signal flow between stages. Our attention module mimics the depth-wise convolution [54] to conduct information aggregation, which has the content-aware property of self-attention while remaining computationally efficient. The most related works to our module are the methods that learn the dynamic filter for restoration [6], [55], [56]. Instead of directly applying the learned attention weights to the input feature, we perform filter modulation in advance to accentuate the informative spectral component in the feature by rescaling the importance of the high-pass filter in the attention map. Furthermore, our module provides multi-shape representation learning and does not produce as many attention weights as them, resulting in fewer parameters and lower complexity.

2.3 Spectral Networks

Since there is a big difference between the spectral features of sharp and degraded image pairs [30], [57], frequency processing is widely adopted in the conventional algorithms for the restoration problem [58], [59]. Recently, researchers have incorporated frequency-based modules into CNNs and Transformer models to bridge the spectral gap. For instance, Mao *et al.* [57] enables low- and high-frequency learning for motion deblurring based on the Fourier transform and CNNs. Zou *et al.* [23] propose a wavelet-based reconstruction module to recover more high-frequency details. Yu *et al.* [60] reconstruct the phase component under the guidance of the amplitude spectrum by revisiting the haze degradations in the frequency domain. Zhou *et al.* [61] incorporate a Fourier-based general prior into the spatial interaction and channel evolution. Chen *et al.* [28] present a hierarchical network for snow removal based on the dual-tree complex wavelet transform [62]. The common practice of these algorithms is first to transform spatial features into the frequency domain through wavelet and Fourier transforms, and then utilize convolutions to modulate the resulting spectra.

Instead of following the above-mentioned paradigm of transform-CNN-inverse transform, ConvIR performs filter modulation on the attention weights using lightweight attention parameters. As such, the importance of filters for informative frequency signals is lifted. Furthermore, our refined filters are imposed on spatial features without transforming these features into the spectral domain using any existing transformation tools, such as Fourier and wavelet transforms, saving computation overhead.

3 Method

In this section, we first describe the overall pipeline of our network. Then, we present the core components: Multi-Scale Module (MSM) and Multi-Shape Attention (MSA). The loss functions are introduced in the final part.

3.1 Overall Architecture

As illustrated in Figure 3 (a), the proposed network adopts a U-shaped architecture for image restoration. Specifically, given any degraded image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, ConvIR first applies a 3×3 convolution layer to generate the shallow features with the size of $C \times H \times W$, where C denotes the number of channels and $H \times W$ represents spatial locations. Then, the shallow features pass through three CNNBlocks to yield the in-depth features. Each CNNBlock contains multiple residual blocks with our MSM inserted into the last one, as depicted in Figure 3 (c). During this process, the channels are expanded, whereas the spatial resolution is reduced. Moreover, following previous algorithms [5], [47], [57], multiple downsampled degraded images are merged into the main path to better handle different blur levels in images. Concretely, ConvS is used to extract the features from the downsampled degraded images by gradually increasing the number of channels. Subsequently, the extracted features are concatenated with those from the main path, followed by a convolution to reduce the channel quantity.

Next, the in-depth features are fed into another three CNNBlocks to restore the high-resolution features. During

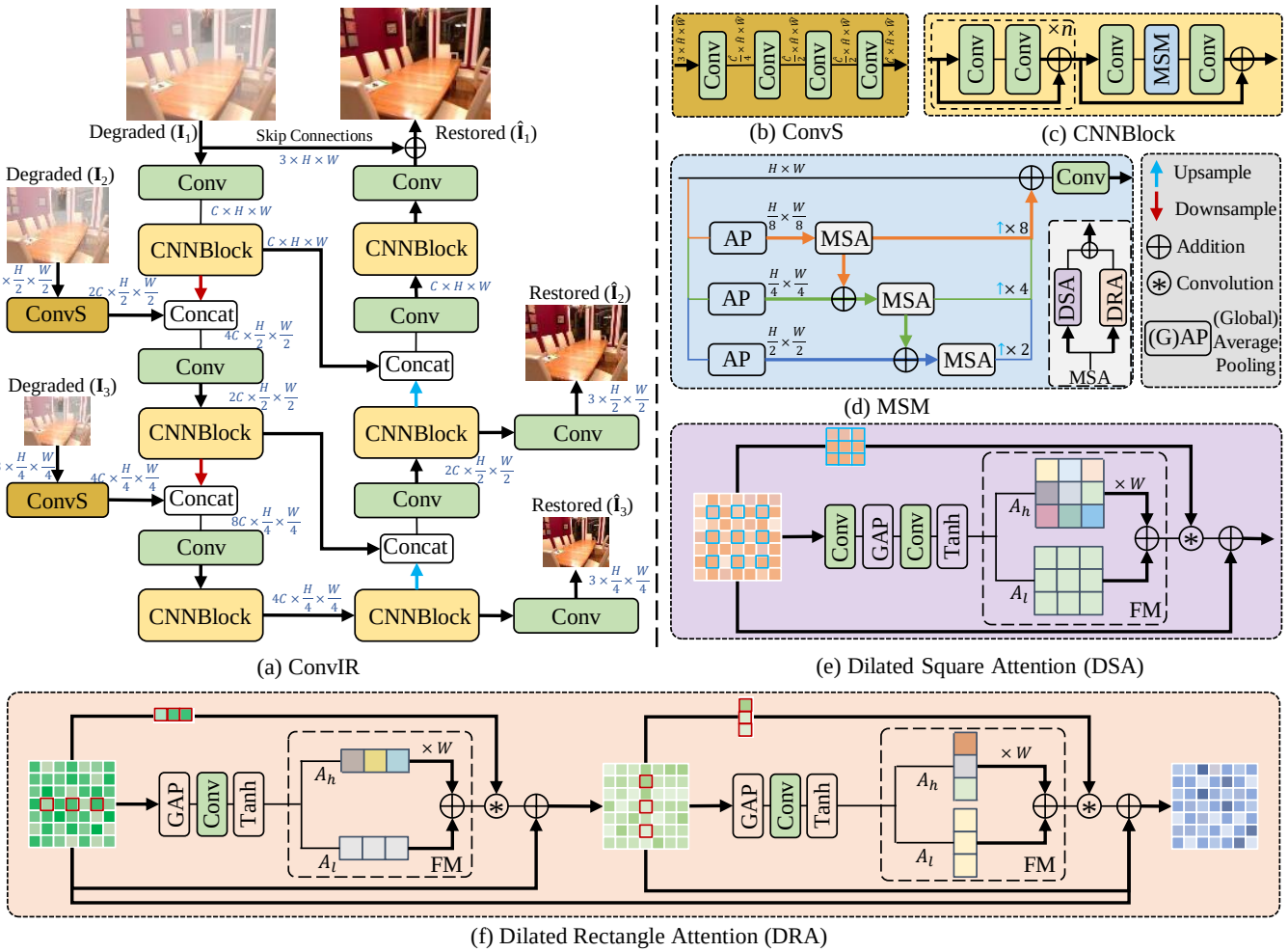


Fig. 3. The architecture of the proposed ConvIR. **(a)** ConvIR comprises six CNNBlocks and adopts the multi-input and multi-output strategies for image restoration. **(b)** ConvS extracts the shallow features from low-resolution degraded images, which includes a series of convolutions with kernel sizes of 3×3, 1×1, 3×3, and 1×1, and gradually increases the channel number from 3 to the target quantity. **(c)** CNNBlock contains multiple residual blocks with the proposed multi-scale module (MSM) inserted into the last one. **(d)** MSM provides multi-scale representation learning in each scale of the U-shaped network. The main operator, multi-shape attention (MSA), contains dilated square attention (DSA) and dilated rectangle attention (DRA) in parallel. **(e)** DSA performs information aggregation within dilated square regions based on filter modulation (FM). **(f)** DRA harvests signals within two-directional rectangles based on FM.

training, the multi-output strategy is adopted, where the low-resolution clean images are predicted after the first two CNNBlocks of the decoder using 3×3 convolutions and image-level skip connections, which are omitted in Figure 3 (a) for simplicity. Furthermore, decoder features are concatenated with the encoder features to assist restoration, and a 1×1 convolution layer is subsequently used to halve the number of channels. The residual clean image is produced via a 3×3 convolution, to which the degraded input is added to output the final restored image. Next, we detail the proposed modules: MSM and MSA.

3.2 Multi-Scale Module (MSM)

The single encoder-decoder paradigm is commonly applied in recent deep restoration architectures to learn hierarchical representations efficiently. However, the number of scales in those works is limited to handle degradation blurs of different sizes. To enhance multi-scale learning and remove blurs in a coarse-to-fine manner in each scale, we mimic the multi-stage network and implement it in each scale of a single U-shaped framework, as illustrated in Figure 1 (d).

The architecture of MSM is shown in Figure 3 (d). For an input tensor $\mathbf{X}\in\mathbb{R}^{H\times W}$, where the channel dimension is ignored for clarity, our MSM utilizes average pooling (AP) operators with different downsampling ratios to convert $\mathbf{X}$ into distinct features spaces. In each branch, the resulting features after MSA are incorporated into the next branch via an addition operator. In this way, MSM can remove degradations progressively by imitating the multi-stage network. Finally, the outputs of all branches are unified to the original input size and added together. In ConvIR, we empirically adopt three branches plus the identity connection, where the downsampling rates are set to {8,4,2}. For the i^{th} ($i \in \{1, 2, 3\}$) branch (except the identity path), the output features can be obtained by:

$$\hat{\mathbf{X}}_i = \mathrm{MSA}(\mathrm{AP}_{2^{4-i}}(\mathbf{X}) + \hat{\mathbf{X}}_{i-1} \uparrow_2) \uparrow_{2^{4-i}}, \quad (1)$$

where $\hat{\mathbf{X}}_0$=$\mathbf{0}$; $\mathrm{AP}_{2^{4-i}}$ denotes average pooling with the downsampling rate as 2^{4-i}; and $\uparrow_2$ represents the *bilinear*

interpolation with the upsampling rate as 2. To summarize, the whole process of MSM can be formally expressed as:

$$\hat{\mathbf{X}} = \mathrm{Conv}_{3\times3}(\sum_{i=1}^{3}\hat{\mathbf{X}}_i + \mathbf{X}), \tag{2}$$

where $\mathrm{Conv}_{3\times3}$ denotes a convolution of 3×3 kernel size.

3.3 Multi-Shape Attention (MSA)

To facilitate multi-scale learning, we aim to devise an efficient module inserted into each branch of MSM to refine features. Equipped with self-attention, Transformer models have achieved promising performance on various image restoration tasks [13], [14]. Despite a few remedies [15], [17], however, the issue of quadratic complexity of self-attention remains intractable. On the other hand, the convolution operator has the static filter, which is incompetent to deal with spatially-varying degradation blurs [16].

In this work, we present MSA by combining the merits of self-attention and convolution operator. Our MSA inherits the content-aware property of the former and maintains the efficiency characteristic of the latter. Furthermore, our MSA involves operators with different shapes and dilation mechanisms with different rates to enhance multi-shape and multi-scale representation learning and enlarge receptive fields. As illustrated in Figure 3 (d), MSA consists of Dilated Square Attention (DSA) and Dilated Rectangle Attention (DRA) in parallel. Next, we delineate DSA and DRA.

3.3.1 Dilated Square Attention (DSA)

As presented in Figure 3 (e), DSA leverages a simple convolution block to generate attention weights, which are adaptive to the input feature, and then performs aggregation using the convolution operation. In the canonical self-attention, Softmax is used to normalize attention weights. However, the resulting sum-to-one weights can be considered as the kernel of a low-pass filter [63], which is unsuitable for image restoration, because the significant discrepancies between the sharp and degraded images mainly lie in the high-frequency components [30], [57].

We resolve the above issue in the attention weights generation step from two aspects: (*i*) bypassing the limitation of low-pass filters with the hyperbolic tangent function (Tanh), and (*ii*) elevating the significance of high-pass filters in attention weights via the proposed filter modulation (FM).

Utilization of Tanh. We substitute Tanh for Softmax. This scheme enjoys two advantages. Firstly, we steer clear of the limitation of the low-pass filter. Secondly, since Tanh projects attention weights into (-1, 1), the negative weights can help suppress the detrimental pixels when performing information aggregation. Formally, given $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the attention weights generating process can be expressed as:

$$\mathbf{A}^{DSA} = \mathrm{Tanh}(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(\mathrm{Conv}_{3\times3}(\mathbf{X})))), \tag{3}$$

where GAP denotes the global average pooling, and Tanh represents the hyperbolic tangent function. To strike a better trade-off between the complexity and diversity of attention weights, instead of producing attention weights for each channel [6], [55], we impose attention weights on the input feature in groups. In each feature group, attention weights are shared across channel and spatial dimensions. $\mathbf{A}^{DSA} \in \mathbb{R}^{G\times K\times K}$, where G is the number of groups and K^2 is the absolute region size for integration.

FM. In addition to using Tanh, we propose reweighing the ratio of high-pass filters in the attention map to enable the network to focus more on the informative frequency components. To this end, as illustrated in Figure 3 (e), we first decompose the attention map $\mathbf{A}^{DSA}$ into low-/high-pass filters, and then reweigh the high-pass one using trainable channel-wise parameters. Thus, the reassembled filter becomes adaptive to emphasize the useful frequency. In practice, due to its ease of implementation, we refer to the low-pass filter as a particular filter that only preserves the direct-current component of the input, which can be extracted from $\mathbf{A}^{DSA}$ by:

$$\mathbf{A}_l^{DSA} = \frac{1}{K^2}\mathbf{E}, \tag{4}$$

where $\mathbf{E}\in \mathbb{R}^{G\times K\times K}$ has the same shape as $\mathbf{A}^{DSA}$ with all values being 1. See Appendix for more details of Eq. 4. Then, the high-pass filter can be considered as the complementary part of the low-pass filter:

$$\mathbf{A}_h^{DSA} = \mathbf{A}^{DSA} - \mathbf{A}_l^{DSA}. \tag{5}$$

Next, the modulated attention map can be obtained by:

$$\tilde{\mathbf{A}}^{DSA} = \mathbf{A}_l^{DSA} + W\mathbf{A}_h^{DSA}, \tag{6}$$

where W denotes the learnable parameters directly optimized by backpropagation and initialized as $\mathbf{1}$.

Finally, we apply the resulting attention weights to the input feature via the convolution operation, where the pixels from the input features are sampled in a dilated manner for a large receptive field. Formally, for each channel in the g^{th} group, the output can be obtained by:

$$\hat{\mathbf{X}}_{g,h,w} = \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\mathbf{X}_{g,(h-\lfloor\frac{K}{2}\rfloor+i)d,(w-\lfloor\frac{K}{2}\rfloor+j)d}\tilde{\mathbf{A}}_{g,i,j}^{DSA} + \mathbf{X}_{g,h,w}, \tag{7}$$

where g, h, w are the indexes of the group, height, and width, respectively. d denotes the dilation rate.

3.3.2 Dilated Rectangle Attention (DRA)

Apart from DSA, we propose DRA that integrates information within rectangles in orthogonal directions to improve multi-shape representation learning. The architecture is shown in Figure 3 (f). Similar to DSA, we employ a convolutional network for generating raw attention weights, the high-pass filter of which is then reassessed through rectangle-shaped FM. Subsequently, the modulated attention weights are imposed on the dilated pixels of input features for information integration. Taking the horizontal unit as an example and denoting the input features as $\mathbf{X}\in \mathbb{R}^{C\times H\times W}$, the output features of each group can be formally obtained by:

$$\hat{\mathbf{X}}_{g,h,w}^{H} = \sum_{j=0}^{K-1}\mathbf{X}_{g,h,(w-\lfloor\frac{K}{2}\rfloor+j)d}\tilde{\mathbf{A}}_{g,j}^{H} + \mathbf{X}_{g,h,w}, \tag{8}$$

$$\tilde{\mathbf{A}}^{H} = \mathrm{RFM}(\mathbf{A}^{H}), \tag{9}$$

$$\mathbf{A}^{H} = \mathrm{Tanh}(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(\mathbf{X}))) \in \mathbb{R}^{G\times K}, \tag{10}$$

where g, h, w index the group, height, and width; d is the dilation rate; A^H denotes the attention map of the

horizontal unit in DRA; RFM means applying FM to the rectangle-shaped attention weights. The horizontal unit can be concluded as $\hat{\mathbf{X}}^H = \mathcal{H}(\mathbf{X})$.

Similarly, taking $\hat{\mathbf{X}}^H$ as the input, the output of the vertical unit can be yielded by $\hat{\mathbf{X}}^V = \mathcal{V}(\hat{\mathbf{X}}^H)$. Finally, the result of DRA is generated by $\hat{\mathbf{X}} = \hat{\mathbf{X}}^V + \mathbf{X}$.

3.4 Loss Functions

Since we introduce FM in MSA, besides the spatial $\mathcal{L}_1$ loss, we adopt the spectral $\mathcal{L}_1$ loss to accentuate the useful frequency. The dual-domain loss functions are given by:

$$\mathcal{L}_{spatial} = \sum_{i=1}^{3} \frac{1}{P_i} \|\hat{\mathbf{I}}_i - \mathbf{Y}_i\|_1, \tag{11}$$

$$\mathcal{L}_{frequency} = \sum_{i=1}^{3} \frac{1}{S_i} \|[\mathcal{R}(\hat{\mathbf{I}}_i), \mathcal{I}(\hat{\mathbf{I}}_i)] - [\mathcal{R}(\mathbf{Y}_i), \mathcal{I}(\mathbf{Y}_i)]\|_1, \tag{12}$$

where i is the index of multiple outputs as shown in Figure 3 (a); $\hat{\mathbf{I}}$ and $\mathbf{Y}$ are the restored image and ground truth, respectively; P and S denote the total elements for normalization; $[\cdot, \cdot]$ is a concatenation operator; and $\mathcal{R}$ and $\mathcal{I}$ are the real and imaginary components yielded by the fast Fourier transform. The final loss function is obtained by:

$$\mathcal{L}_{total} = \mathcal{L}_{spatial} + \lambda \mathcal{L}_{frequency}, \tag{13}$$

where λ is set to 0.1 for balancing dual-domain training.

TABLE 1
Image dehazing comparisons on the synthetic SOTS [27] dataset.

Methods	SOTS-Indoor PSNR	SOTS-Indoor SSIM	SOTS-Outdoor PSNR	SOTS-Outdoor SSIM	Params (M)	FLOPs (G)
GridDehazeNet [20]	32.16	0.984	30.86	0.982	0.956	21.5
MSBDN [10]	33.67	0.985	33.48	0.982	31.35	41.54
FFA-Net [11]	36.39	0.989	33.57	0.984	4.456	287.8
AECR-Net [68]	37.17	0.990		-	2.611	52.2
DeHamer [49]	36.63	0.988	35.18	0.986	132.50	60.3
DehazeFormer-L [51]	40.05	<u>0.996</u>		-	25.44	279.7
MAXIM [47]	38.11	0.991	34.19	0.985	14.1	216
PMNet [69]	38.41	0.990	34.74	0.985	18.90	81.13
MB-TaylorFormer-B [31]	40.71	0.992	37.42	0.989	2.68	38.5
MB-TaylorFormer-L [31]	<u>42.64</u>	0.994	<u>38.09</u>	0.991	7.43	88.1
ConvIR-S (Ours)	41.53	<u>0.996</u>	37.95	<u>0.994</u>	5.53	42.1
ConvIR-B (Ours)	**42.72**	**0.997**	**39.42**	**0.996**	8.63	71.22

TABLE 2
Image dehazing results on the Haze4K [38] dataset.

Methods	PSNR	SSIM	Params/M	FLOPs/G
DehazeNet [43]	19.12	0.84	0.01	0.58
AOD-Net [70]	17.15	0.83	0.002	0.12
GridDehazeNet [20]	23.29	0.93	0.956	21.5
MSBDN [10]	22.99	0.85	31.35	41.54
FFA-Net [11]	26.96	0.95	4.456	287.8
DMT-Net [38]	28.53	0.96	-	-
PMNet [69]	33.49	<u>0.98</u>	18.90	81.13
FSNet [71]	34.12	**0.99**	13.28	110.5
ConvIR-S (Ours)	33.36	**0.99**	5.53	42.1
ConvIR-B (Ours)	<u>34.15</u>	**0.99**	8.63	71.22
ConvIR-L (Ours)	**34.50**	**0.99**	14.83	129.34

4 Experiments

To verify the efficacy of our method, we evaluate ConvIR on 20 different datasets for five image restoration tasks: image dehazing, image defocus deblurring, image desnowing, image deraining, and image motion deblurring. In this section, we first introduce the experimental setup. Then, we present the results of our models and compare them quantitatively and qualitatively with state-of-the-art schemes. Finally, we conduct extensive ablation studies to verify the efficacy of our proposed components. In the tables, the best and second best results are in **boldface** and <u>underlined</u>, respectively.

4.1 Experimental Setup

Implementation details. We train separate models for different problems. Unless mentioned otherwise, the following hyper-parameters are adopted in all experiments. The number of the group (G) and the region size (K) are set to 8 and 3, respectively. The dilation rates in the three branches (from top to bottom) of MSM are 7, 9, and 11, respectively. We train our model using the Adam optimizer [64] with the initial learning rate as $1e^{-4}$, which is gradually reduced to $1e^{-6}$ with cosine annealing [65]. For data augmentation, we only use random horizontal flips. According to the complexity of different problems, we introduce three ConvIR variants in our experiments by varying the number of regular residual blocks (n) in CNNBlock to validate the effectiveness comprehensively. Expressly, we set $n = 3$, $n = 7$, and $n = 15$ in our ConvIR-S (Small), ConvIR-B (Base), and ConvIR-L (Large). All models are trained on an NVIDIA Tesla A100 GPU with PyTorch. More details of the datasets and specific training configurations are provided in the Appendix.

Evaluation metrics. We adopt the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [66] metrics for all datasets. Higher PSNR and SSIM indicate that the restored image is closer to the target one in terms of pixel-wise contents and structural properties. In addition, mean absolute error (MAE) and learned perceptual image patch similarity (LPIPS) [67] are employed for defocus deblurring, with a lower score indicating better performance. Unless stated otherwise, FLOPs are measured on 256×256 patches.

4.2 Experimental Results

4.2.1 *Image Dehazing*

For this problem, we first compare our methods with state-of-the-art algorithms on the SOTS [27] dataset in Table 1. As can be seen, our ConvIR-B achieves the best accuracy results on all metrics. In particular, ConvIR-B outperforms the first Transformer-based dehazing algorithm, DeHamer [49], by 6.09 dB and 4.24 dB PSNR on the SOTS-Indoor [27] and SOTS-Outdoor [27] datasets, respectively, with 93% fewer parameters. Our small model is significantly superior to the expensive Transformer-based DehazeFormer-L [51] with a performance gain of 1.48 dB PSNR on SOTS-Indoor, consuming 85% lower complexity. Furthermore, our two models surpass the corresponding variants of the recent MB-TaylorFormer [31] with comparable computation overhead. Moreover, we present the comparisons on a more realistically synthetic Haze4K [38] dataset in Table 2. ConvIR-B

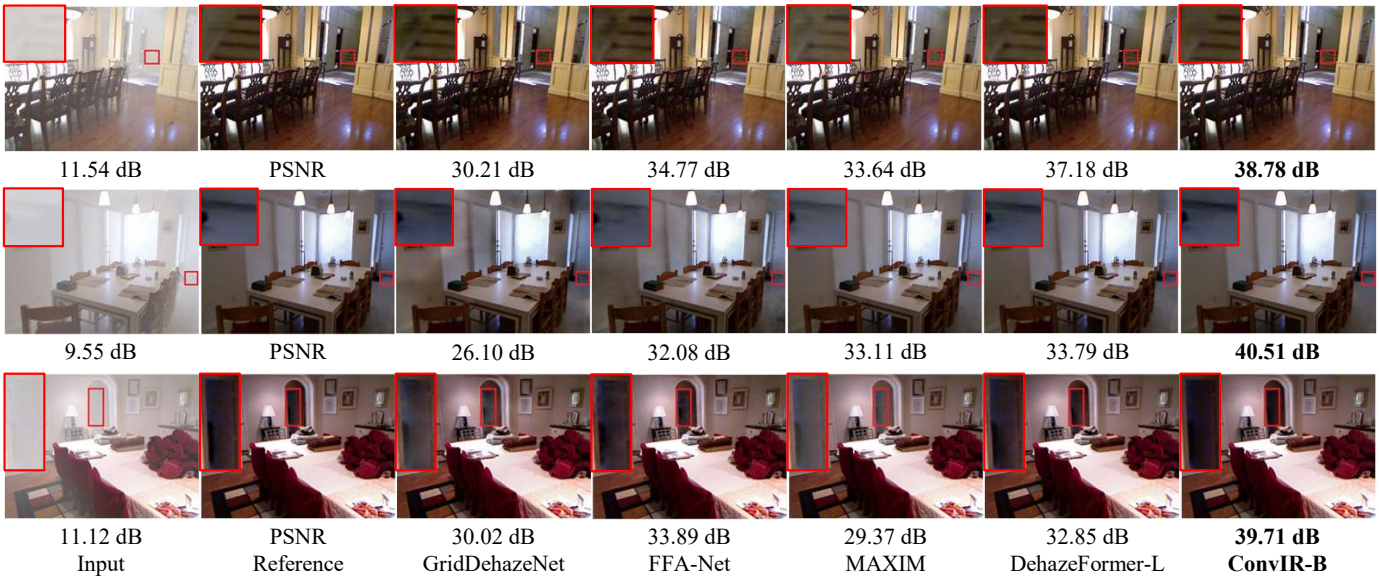


Fig. 4. Image dehazing comparisons on the SOTS-Indoor [27] dataset.

Fig. 5. Image dehazing comparisons on the SOTS-Outdoor [27] dataset.

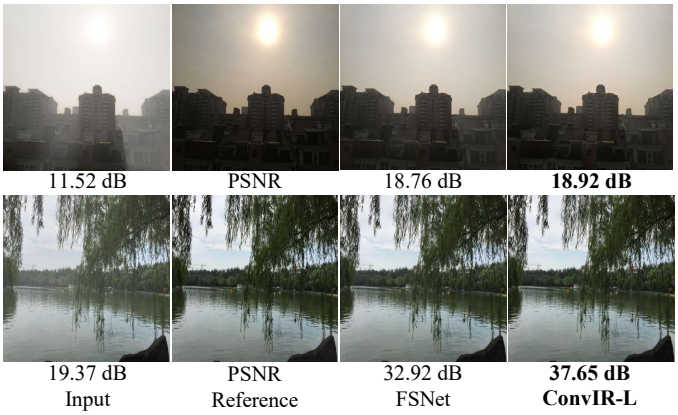


Fig. 6. Image dehazing comparisons on the Haze4K [38] dataset.

outperforms the recent FSNet [71] by 0.03 dB PSNR, with only 65% parameters and 64% FLOPs. To further demonstrate the superiority of our model, we present the results of our large model, which has comparable computation costs with FSNet. ConvIR-L produces a substantial performance gain of 0.38 dB PSNR over FSNet [71].

TABLE 3
Image dehazing comparisons on four real-world datasets: Dense-Haze [72], NH-HAZE [73], O-HAZE [39], and I-Haze [40].

Methods	Dense-Haze		NH-HAZE		O-HAZE		I-Haze	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
GridDehazeNet [20]	13.31	0.368	13.80	0.537	18.92	0.672	18.73	0.769
MSBDN [10]	15.13	0.555	19.23	0.706	24.36	0.749	19.62	0.618
FFA-Net [11]	15.70	0.549	19.87	0.692	22.12	0.770	19.72	0.733
DeHamer [49]	16.62	0.560	**20.66**	0.684	25.11	0.777	-	
PMNet [69]	16.79	0.510	20.42	0.730	-		-	
MB-TaylorFormer-B [31]	16.66	0.560	-		25.05	**0.788**	-	
MB-TaylorFormer-L [31]	16.64	0.566	-		25.31	0.782	-	
ConvIR-S (Ours)	**17.45**	**0.648**	20.65	**0.807**	25.25	0.784	21.95	**0.888**
ConvIR-B (Ours)	16.86	0.621	**20.66**	0.802	**25.36**	0.780	**22.44**	0.887

The visual comparisons on these synthetic daytime datasets, SOTS-Indoor [27], SOTS-Outdoor [27], and Haze4K [38], are illustrated in Figure 4, Figure 5, and Figure 6, respectively. The haze-free images generated by our models are more visually faithful to ground-truth images.

We further extensively compare our models with state-of-the-art schemes on four real-world datasets, *i.e.*, Dense-Haze [72], NH-HAZE [73], O-HAZE [39], and I-Haze [40]. Table 3 shows that the best results are mostly generated by

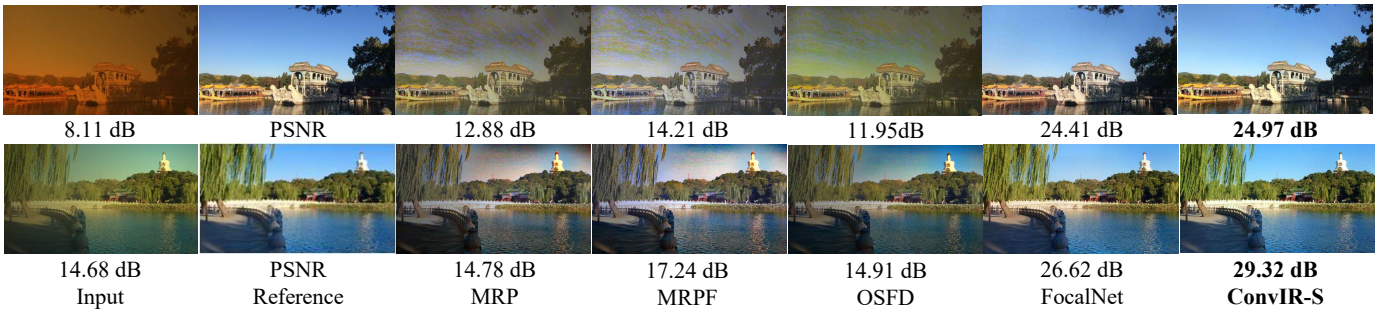


Fig. 7. Nighttime image dehazing comparisons on the NHR [35] dataset.

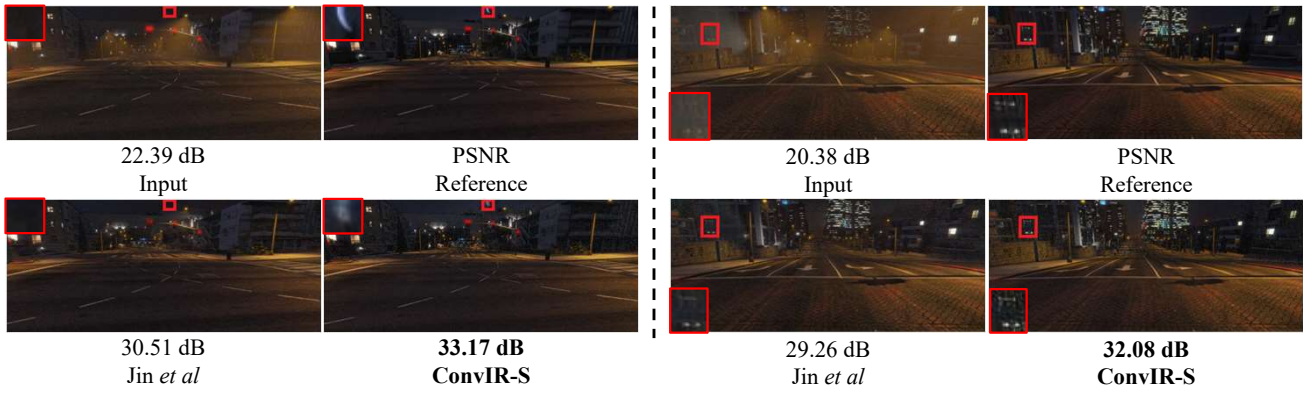


Fig. 8. Nighttime image dehazing comparisons on the GTA5 [36] dataset.

TABLE 4
Image dehazing comparisons on the remote sensing datasets: SateHaze1k-Thin, SateHaze1k-Moderate, and SateHaze1k-Thick [37]. † denotes methods that are specially designed for remote sensing.

Methods	Thin		Moderate		Thick	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
AOD-Net [70]	19.54	0.854	20.10	0.885	15.92	0.731
H2RL-Net† [74]	20.91	0.880	22.34	0.906	17.41	0.768
FCFT-Net† [75]	23.59	0.913	22.88	0.927	20.03	0.816
Uformer [15]	22.82	0.907	24.47	0.939	20.36	0.815
C^2PNet [76]	19.62	0.880	24.79	0.940	16.83	0.790
Restormer [16]	23.08	0.912	24.73	0.933	18.58	0.762
Trinity-Net† [77]	21.55	0.884	23.35	0.895	20.97	0.823
UMWTransformer [78]	24.29	0.919	26.65	0.946	20.07	0.825
FocalNet [46]	24.16	0.916	25.99	0.947	21.69	0.847
ConvIR-S (Ours)	**25.11**	**0.978**	**26.79**	**0.978**	**22.65**	**0.950**

TABLE 5
Nighttime image dehazing comparisons on the NHR [35] dataset. † denotes methods that are specially designed for nighttime dehazing.

Methods	PSNR	SSIM
NDIM† [80]	14.31	0.526
GS† [81]	17.32	0.629
MRPF† [82]	16.95	0.667
MRP† [82]	19.93	0.777
OSFD† [35]	21.32	0.804
HCD [83]	23.43	0.953
FocalNet [46]	25.35	0.969
Jin *et al*† [79]	26.56	0.890
ConvIR-S (Ours)	28.85	0.981
ConvIR-B (Ours)	**29.49**	**0.983**

our models. Specifically, our base version ConvIR-B yields performance gains of 0.22 dB and 0.05 dB PSNR over the recent MB-TaylorFormer-L [31] on the Dense-Haze [72] and O-HAZE [39] datasets, respectively. Noticeably, ConvIR-S outperforms the MB-TaylorFormer-L with a remarkable performance gain of 0.81 dB PSNR on the Dense-Haze with only half complexity.

Since image dehazing plays an essential role in remote sensing. We provide experimental results on the remote sensing SateHaze1k [37] dataset. The models are trained and tested separately on its three sub-sets. Table 4 shows that our model performs best on all metrics. In particular, our small model remarkably outperforms the general image restoration method [46] and remote sensing method [77] by 0.96 dB and 1.68 dB PSNR for the thick level, respectively.

Additionally, we conduct experiments on two nighttime datasets, *i.e.*, NHR [35] and GTA5 [36]. The quantitative comparisons on NHR [35] are presented in Table 5. Our ConvIR-B and ConvIR-S obtain the best and second-best results, respectively. In particular, ConvIR-S outperforms the recent algorithm [79] with a remarkable gain of 2.29 dB PSNR, using 3.8× fewer parameters. Figure 7 illustrates that our ConvIR-S restores a crisper daytime image than other algorithms. Moreover, we provide results on another nighttime dehazing dataset, GTA5, whose ground-truth images are in the nighttime scenes. Table 6 shows that our two versions are superior to the algorithm [79], which is specially devised for nighttime haze removal. Figure 8 demonstrates that our model is robust in nighttime scenes.

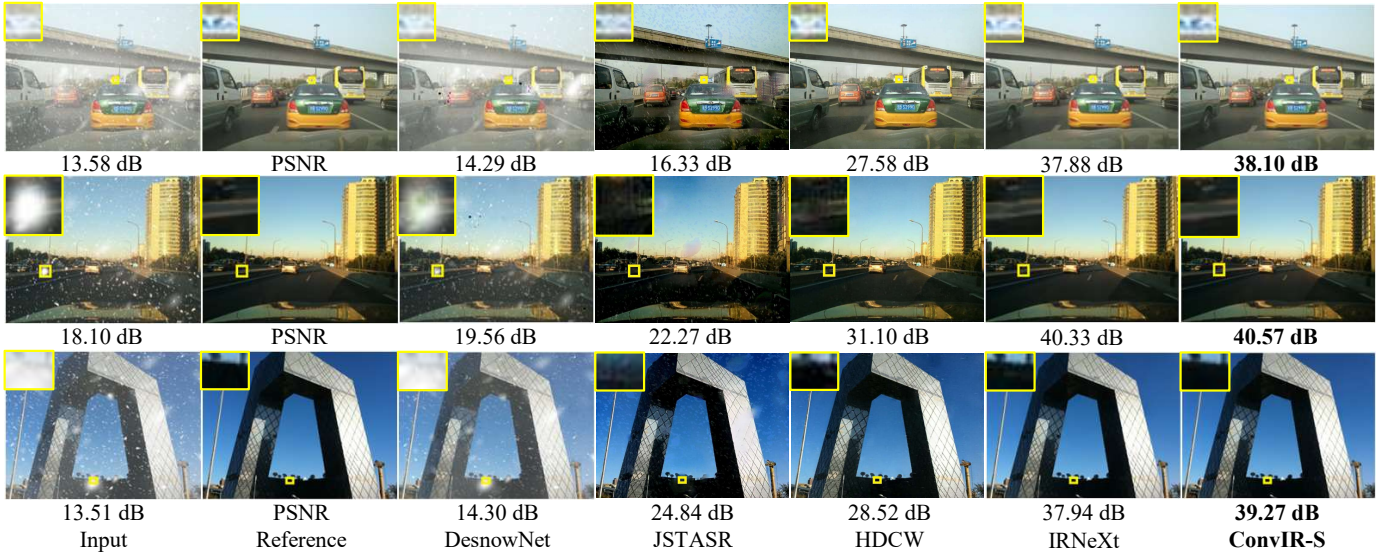


Fig. 9. Image desnowing comparisons on the CSD [28] dataset.

TABLE 6
Nighttime image dehazing comparisons on the GTA5 [36] dataset. † denotes methods that are specially designed for nighttime dehazing.

Methods	PSNR	SSIM
GS† [81]	21.02	0.639
MRP† [82]	20.92	0.646
Ancuti *et al*† [84]	20.59	0.623
Yan *et al*† [36]	27.00	0.850
CycleGAN [85]	21.75	0.696
Jin *et al*† [79]	30.38	0.904
ConvIR-S (Ours)	31.68	0.917
ConvIR-B (Ours)	**31.83**	**0.921**

TABLE 7
Image desnowing comparisons on the CSD [28], SRRS [86], and Snow100K [87] datasets.

Methods	CSD		SRRS		Snow100K		Params (M)	FLOPs (G)
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
DesnowNet [87]	20.13	0.81	20.38	0.84	30.50	0.94	15.6	1.7K
All in One [88]	26.31	0.87	24.98	0.88	26.07	0.88	44	12.26
JSTASR [86]	27.96	0.88	25.82	0.89	23.12	0.86	65	-
HDCW-Net [28]	29.06	0.91	27.78	0.92	31.54	0.95	6.99	9.78
SMGARN [89]	31.93	0.95	29.14	0.94	31.92	0.93	6.86	450.3
TransWeather [33]	31.76	0.93	28.29	0.92	31.82	0.93	21.9	5.64
MSP-Former [50]	33.75	0.96	30.76	0.95	33.43	**0.96**	2.83	4.42
FocalNet [46]	37.18	**0.99**	31.34	**0.98**	33.53	0.95	3.74	30.63
IRNeXt [34]	37.29	**0.99**	31.91	**0.98**	33.61	0.95	5.46	42.09
ConvIR-S (Ours)	38.43	**0.99**	32.25	**0.98**	33.79	0.95	5.53	42.1
ConvIR-B (Ours)	**39.10**	**0.99**	**32.39**	**0.98**	**33.92**	**0.96**	8.63	71.22

4.2.2 *Image Desnowing*

We compare desnowing performance on three widely-adopted datasets: CSD [28], SRRS [86], and Snow100K [87]. Table 7 shows that our ConvIR-B yields the best results on all metrics. In particular, ConvIR-B outperforms the recent IRNeXt [34] by 0.48 dB and 0.31 dB in terms of PSNR on SRRS and Snow100K, respectively, with comparable computation overhead. On a challenging CSD dataset containing more intricate snow scenes, the advantage of our ConvIR-B becomes more pronounced, showcasing the superior ability of our network in snow removal. It is worth mentioning that our small model is superior to the Transformer-based TransWeather [33] on all metrics while consuming 75% fewer parameters. Furthermore, compared with MSP-Former [50], which is elaborately designed for desnowing, our small model shows a significant performance boost of 4.68 dB PSNR on the CSD dataset.

Figure 9 shows that our model is more effective than the competitors in removing snow degradations and recovers more detailed contours without noticeable artifacts, such as the road divider in the second image.

TABLE 8
Image deraining comparisons on Test100 [90] and Test2800 [91].

Methods	Test100		Test2800	
	PSNR	SSIM	PSNR	SSIM
DerainNet [92]	22.77	0.810	24.31	0.861
SEMI [93]	22.35	0.788	24.43	0.782
UMRL [94]	24.41	0.829	29.97	0.905
RESCAN [95]	25.00	0.835	31.29	0.904
PreNet [96]	24.81	0.851	31.75	0.916
MSPFN [97]	27.50	0.876	32.82	0.930
MPRNet [12]	30.27	0.897	33.64	**0.938**
FSNet [71]	31.05	**0.919**	33.64	0.936
ConvIR-L (Ours)	**31.40**	**0.919**	**33.73**	0.937

4.2.3 *Image Deraining*

We perform experiments for image deraining by training the model on a compound dataset [91], [99]–[101]. The evaluation results on Test100 [90] and Test2800 [91] are presented in Table 8. Our model significantly surpasses the CNN-based FSNet [71] and MPRNet [12] by 0.35 dB and 1.13 dB PSNR, respectively, on the Test100 [90] dataset. The superiority of our model can also be found on the Test2800 [91] dataset when compared with other state-of-the-art schemes.

The visual comparisons on the Test100 [90] dataset are illustrated in Figure 10. As we can see, our method generates

Fig. 10. Image deraining comparisons on the Test100 [90] dataset.

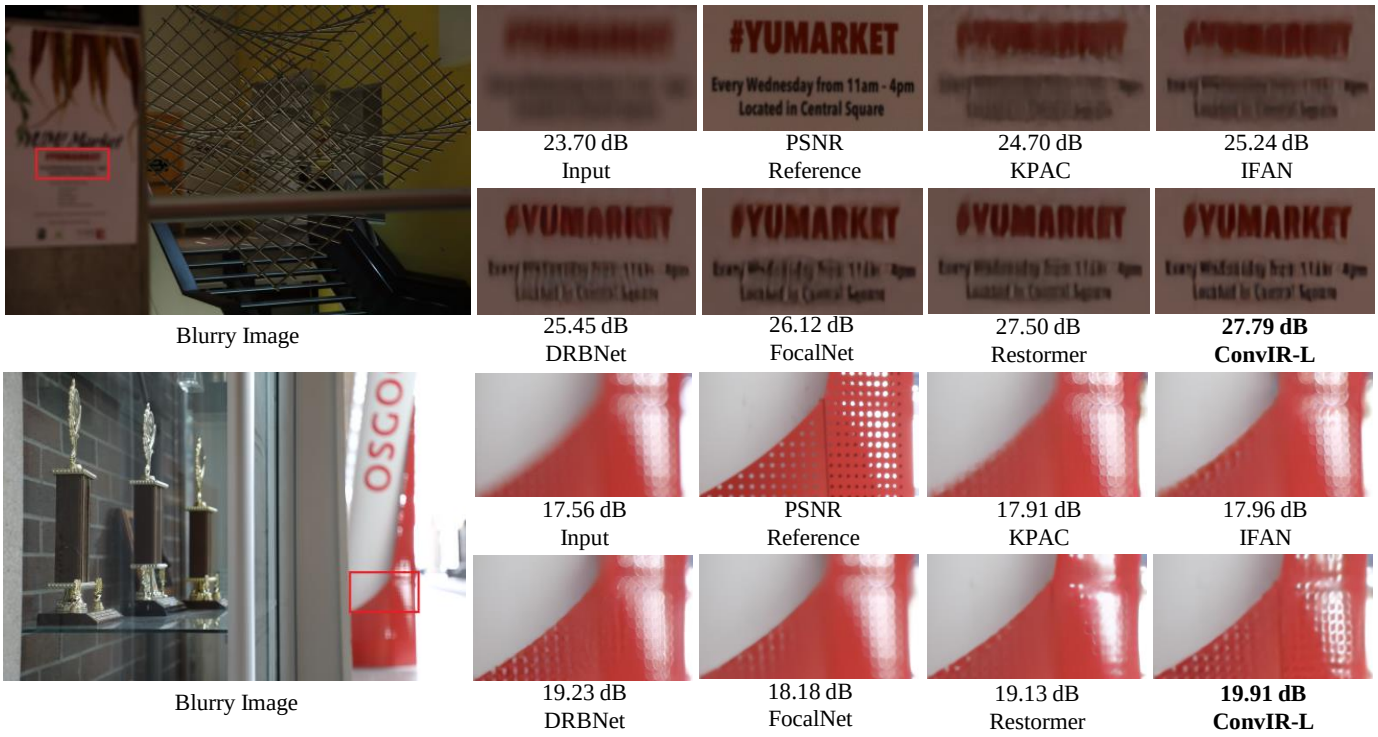


Fig. 11. Image defocus deblurring comparisons on the DPDD [29] dataset.

TABLE 9
Image defocus deblurring comparisons on the DPDD [29] dataset. FLOPs are measured under the resolution of $3 \times 720 \times 1280$.

Methods	Indoor Scenes				Outdoor Scenes				Combined				Params (M)	FLOPs (G)
	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓		
DPDNet [29]	26.54	0.816	0.031	0.239	22.25	0.682	0.056	0.313	24.34	0.747	0.044	0.277	31.03	770
KPAC [7]	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227	2.06	113
MDP [98]	28.02	0.841	0.027	-	22.82	0.690	0.052	-	25.35	0.763	0.040	-	46.86	1898
IFAN [6]	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217	10.48	363
DRBNet [19]			-				-		25.73	0.791	-	0.183	11.69	693
FocalNet [46]	29.10	0.876	0.024	0.173	23.41	0.743	**0.049**	0.246	26.18	0.808	0.037	0.210	12.82	1376
Restormer [16]	28.87	0.882	0.025	0.145	23.24	0.743	0.050	0.209	25.98	0.811	0.038	0.178	26.16	1983
IRNeXt [34]	29.22	0.879	0.024	0.167	**23.53**	0.752	**0.049**	0.244	26.30	0.814	0.037	0.206	14.76	1778
ConvIR-S (Ours)	28.95	0.877	0.024	0.158	23.32	0.747	0.050	0.221	26.06	0.810	0.037	0.190	5.53	579
ConvIR-B (Ours)	29.06	0.879	0.024	0.156	23.42	0.752	**0.049**	0.219	26.16	0.814	0.037	0.188	8.63	979
ConvIR-L (Ours)	**29.37**	**0.887**	**0.023**	**0.143**	23.51	**0.757**	0.049	**0.203**	**26.36**	**0.820**	**0.036**	**0.174**	14.83	1778

a higher quality image than other competitors by better removing rainy degradations and restoring color.

4.2.4 *Image Defocus Deblurring*

We conduct image defocus deblurring experiments on the widely used DPDD [29] dataset with our three variants to comprehensively compare with state-of-the-art approaches. The image fidelity scores are presented in Table 9. Our ConvIR-L achieves 26.36 dB PSNR in the combined category, which is 0.06 dB higher than the recent CNN-based IRNeXt [34] algorithm with similar parameters and FLOPs. Compared to the strong Transformer-based Restormer [16], our large model obtains a substantial average performance gain of 0.38 dB in terms of PSNR. It is worth noting that

Fig. 12. Image motion deblurring comparisons on the GoPro [25] dataset.

TABLE 10
Image motion deblurring results on GoPro [25]. The inference time is tested on $3 \times 720 \times 1280$ images in a synchronized manner with *torch.cuda.synchronize*() using an NVIDIA Tesla V100 GPU.

Methods	PSNR	SSIM	FLOPs (G)	Params (M)	Time (s)	Memory (G)
MIMO-UNet++ [5]	32.68	0.959	617.64	16.1	1.277	10.395
MPRNet [12]	32.66	0.959	777.01	20.1	1.148	10.415
MAXIM-3S [47]	32.86	0.961	119.5	22.2	-	-
Restormer [16]	32.92	0.961	140.99	26.1	1.218	12.333
Stripformer [17]	33.08	0.962	170.46	20.0	1.054	12.149
PromptRestorer [32]	33.06	0.962	-	-	-	-
IRNeXt [34]	33.19	**0.963**	129.33	14.76	0.291	6.865
ConvIR-L (Ours)	**33.28**	**0.963**	129.34	14.83	0.323	6.867

our small version still outperforms Restormer with a gain of 0.08 dB in PSNR for the combined scenes despite using only 21% parameters and 29% FLOPs.

The visual comparisons are illustrated in Figure 11. As seen, our method recovers more structural details from hard defocus degradations, such as the words on the poster.

4.2.5 Image Motion Deblurring

We evaluate our model on a widely used synthetic GoPro [25] dataset and a real-world RSBlur [102] dataset. The overall comparisons in terms of accuracy and computational costs on the GoPro [25] dataset are presented in Table 10. Compared with Transformer models Restormer [16] and Stripformer [17] that have quadrative complexity, our network, ConvIR-L, is built on the efficient convolutional network and receives remarkable performance gains of 0.36 dB and 0.20 dB PSNR respectively, with fewer parameters, lower complexity, and less memory footprint. Furthermore, our model runs 3.77× and 3.26× faster than these two

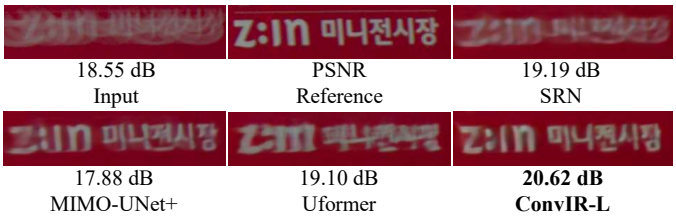


Fig. 13. Image deblurring comparisons on the real-world RSBlur [102].

algorithms, respectively, suggesting that our model strikes a better trade-off between the accuracy and computation overhead. Compared to the more recent Transformer model, PromptRestorer [32], our method continues to demonstrate superior performance, achieving a 0.22 dB higher score. Compared to CNN-based methods, such as MPRNet [12] and MIMO-UNet++ [5], our method showcases superiority on all metrics due to the more advanced network design, such as the cost-effective frequency modules. Additionally, our model surpasses IRNeXt [34] by 0.09 dB PSNR with similar computation overhead owing to the cheap two-directional rectangle attention unit and dilated operation.

The visual comparisons on the GoPro [25] dataset are illustrated in Figure 12. Our ConvIR-L produces more visually pleasing results than other algorithms by removing large motion blurs.

In addition, we report the results on the real-world RSBlur [102] in Table 11. As can be seen, our method outperforms the Transformer-based Uformer [15] and Restormer [16] by 0.08 dB and 0.37 dB in PSNR, respectively. The visual comparisons in Figure 13 illustrate that the image produced by our model is much closer to the reference, demonstrating the robust property of our method in real-world scenarios.

TABLE 11
Image motion deblurring results on the real-world RSBlur [102] dataset.

Methods	PSNR	SSIM
SRN-DeblurNet [103]	32.53	0.840
MIMO-UNet [5]	32.73	0.846
MIMO-UNet+ [5]	33.37	0.856
MPRNet [12]	33.61	0.861
Restormer [16]	33.69	0.863
Uformer [15]	33.98	0.866
ConvIR-L (Ours)	**34.06**	**0.868**

TABLE 12
Break-down ablation studies toward better performance. To separately study the effect of MSM, we deploy a 3×3 convolution in each branch to form MSM/Conv. Here, MSM denotes the pure multi-scale paradigm without MSA. DSA/Conv denotes a degraded version of DSA by excluding the FM and dilation mechanism.

Methods	a	b	c	d	e	f
Baseline	✓	✓	✓	✓	✓	✓
MSM/Conv		✓	✓	✓	✓	✓
DSA/Conv			✓	✓	✓	✓
DSA/FM				✓	✓	✓
DSA/Dilation					✓	✓
DRA						✓
PSNR	31.23	31.46	31.53	31.64	31.76	31.92
Params/M	6.90	8.45	8.55	8.56	8.56	8.63
FLOPs/G	66.32	71.17	71.19	71.19	71.19	71.22
Time/s	0.134	0.152	0.165	0.166	0.170	0.206

4.3 Ablation Study

Following previous schemes [34], [47], we conduct ablation studies on the GoPro [25] dataset with ConvIR-B. The baseline model is obtained by removing MSM and its inclusions from our network. All models are trained for 1000 epochs.

Break-down ablation. We perform the break-down ablations by applying our components to the baseline successively. The results are reported in Table 12. The baseline receives 31.23 dB PSNR on GoPro (Table 12a). After deploying MSM with only a 3×3 convolution in each branch, the model achieves a 0.23 dB improvement (Table 12b). We then replace this 3×3 convolution with DSA/Conv, a degraded version of DSA by excluding FM and the dilation mechanism, the model obtains a further boosted performance of 0.07 dB PSNR (Table 12c). Refining filters in the attention map using FM leads to a performance gain of 0.11 dB PSNR (Table 12d). Enlarging receptive fields with the dilation mechanism advances the performance to 31.76 dB PSNR (Table 12e). Additionally deploying rectangle-shaped attention, the complete model achieves the best performance, 0.69 dB higher than the baseline model, and introduces only 1.73M parameters and 4.9G FLOPs. The results suggest the effectiveness of our proposed components.

The number of branches in MSM. The number of branches plays an essential role in the coarse-to-fine mechanism of MSM. Therefore, we conduct experiments by varying the number of branches. Table 13 shows that employing more branches leads to better performance. Specifically, when using a single branch with a downsampling rate of 2, the model receives a gain of 0.22 dB PSNR over the baseline model. When equipped with three branches, the model demonstrates the efficacy of the proposed coarse-to-fine mechanism by producing a 0.69 dB PSNR improvement.

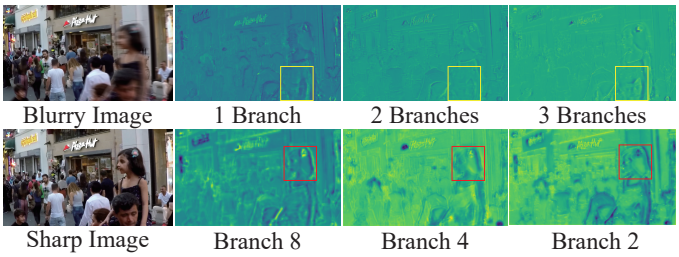


Fig. 14. Visualization of intermediate feature maps. The top three features exhibit the difference in using different numbers of branches in MSM. The corresponding models are shown in Table 13. The feature maps in the second line showcase the outcomes of branches with different downsampling rates when using three branches. The numbers in the bottom line indicate downsampling rates.

TABLE 13
The number of branches in MSM. The number indicates the downsampling rate of a branch.

2	4	8	PSNR	Params/M	FLOPs/G
			31.23	6.90	66.32
✓			31.45	7.74	70.91
✓	✓		31.62	8.18	71.15
✓	✓	✓	31.92	8.63	71.22

TABLE 14
Ablation study for different pooling operations.

Pooling Types	Convolution	Max Pooling	Average Pooling
PSNR	31.75	31.65	31.92

We further visualize the intermediate features in Figure 14 to investigate the effect of our coarse-to-fine mechanism. We first exhibit the features obtained from the last scale of the models that deploy different branches. The features in the first line illustrate that using more branches recovers more high-frequency signals for motion deblurring, such as the outlines in the yellow rectangles. Then, we plot feature maps obtained from each branch of the model that utilizes three branches. The second line shows that the model restores shaper features progressively, such as the face in the red regions. The visualizations demonstrate the effectiveness of the coarse-to-fine mechanism.

Pooling operation choices. We study the influence of using different pooling techniques in MSM, *i.e.,* depth-wise convolution, max pooling, and average pooling. We adopt the same downsampling rate in all variants. These three operations have the same computational complexity, whereas convolution introduces extra parameters. The results are shown in Table 14. The average pooling variant achieves a better result than the other two alternatives. Therefore, we choose average pooling as the default configuration.

Different activation functions. Instead of inheriting Softmax from self-attention to normalize attention weights, we employ Tanh in MSA. This allows the network to steer clear of the low-pass filters and generate negative weights

TABLE 15
Ablation studies for different functions.

Methods	Softmax	Linear	Sigmoid	Tanh
PSNR	31.75	31.79	31.83	31.92

TABLE 16
Ablation studies for the number of groups in MSA.

Methods	2	4	8	16
PSNR	31.80	31.79	31.92	31.85
Params/M	8.51	8.55	8.63	8.79

TABLE 17
Ablation studies for dilation rates. (a, b, c) means setting dilation rates of a, b, and c in three branches of MSM, respectively.

Dilation Rates	(3,2,1)	(7,5,3)	(7,9,11)
PSNR	31.67	31.75	31.76

TABLE 18
Ablation studies for the multi-input and multi-output mechanisms.

Methods	a	b	c	d
Multi-output	✓			
Multi-input (3nd scale)	✓	✓		
Multi-input (2rd scale)	✓	✓	✓	
PSNR	31.23	30.77	30.39	28.97

for pixels that may have a detrimental impact during information aggregation. Table 15 shows that compared with Softmax, the linear projection and the Sigmoid version achieve gains of 0.04 dB and 0.08 dB PSNR, respectively, by breaking away from the sum-to-one property. Tanh projects attention weights into (-1, 1), producing a considerable improvement of 0.17 dB PSNR over the baseline model.

The number of groups in MSA. In MSA, we learn group-wise attention weights for information aggregation. To study the impact of the diversity of attention weights, we perform experiments by varying the number of groups. The results are presented in Table 16. Generally, as we increase the number of groups, the performance improves. However, group 8 appears to be saturated, a phenomenon likely attributed to overfitting.

Dilation rates. We use the dilation mechanism in MSA to enlarge receptive fields. We conduct experiments by deploying different combinations of dilation rates in three branches of MSM. Table 17 shows that increasing the dilation rates leads to better results by perceiving larger receptive fields. Finally, we simply choose the combination of (7,9,11) in our models. Here, we only experiment with a limited set of dilation rate combinations to verify the validity of our design instead of exhaustively searching for the best option.

Effects of multi-input and multi-output strategies. We investigate the effects of the adopted multi-input and multi-output strategies [5], [34], [47] by gradually removing these techniques from the baseline model in Table 12. Table 18 shows that the model receives 31.23 dB PSNR using the multi-input/output mechanisms. Removing the multi-output method, the performance degrades to 30.77 dB PSNR. Unloading multi-input layers leads to further performance degradation. These results suggest the effectiveness of the multi-input and multi-output schemes.

TABLE 19
The overall comparisons between the convolution and self-attention models using the same number of parameters (0.09M). The *memory* is the memory usage for training.

Methods	FLOPs/G	Memory/G	PSNR	Time/s
Conv	6.19	14.3	25.86	0.019
Attention	8.28 (+2.09)	43.7 (+29.4)	25.85 (−0.01)	0.085 (+0.066)

5 Discussion

Recent years have witnessed a paradigm shift from CNN-based architectures to Transformer models, which feature quadratic complexity. Some literature has investigated the connections between the convolutions and Transformer models from various perspectives, including channel mixing, normalization, filter generation [54] and application [104], and frequency preference [63]. In this study, we revitalize the convolutional network simply due to the high complexity of self-attention.

We conduct toy experiments for clear demonstration. Specifically, we build two tiny models by respectively deploying only five residual convolution blocks and five pure self-attention units [105]. We keep the number of parameters equal for a fair comparison by adjusting the channel count. The models are trained on the GoPro [25] dataset for 100 epochs with an initial learning rate of $16e^{-4}$ and a batch size of 64. The obtained models are tested on the GoPro [25] test set using an NVIDIA Tesla V100 GPU. Table 19 shows that the model built on self-attention consumes higher complexity and memory footprint during training. By contrast, the convolution version achieves comparable accuracy with lower computation overhead and faster speed. Therefore, we revitalize the convolution network for effective and efficient image restoration. The comprehensive experimental results demonstrate that using proper designs, the convolutional networks can perform better or favorably against the elaborately devised Transformer models. We hope this study could inspire researchers to further exploit the potential of CNN-based models for image restoration.

6 Conclusion

In this study, we analyze previous successful image restoration models and identify the good properties owned by them. Based on the observation, we present an effective and efficient convolutional model for image restoration. Extensive experimental results on 20 benchmark datasets demonstrate that the proposed network matches Transformer models and achieves state-of-the-art performance for five representative image restoration tasks.

Our work also has limitations. For example, we only experiment with limited combinations of dilation rates in MSM to demonstrate the validity of our design. Promising directions include learning the optimal dilation rates or

combining the deformable operator [106] to capture adaptive and flexible receptive fields. Further work can also involve using cheaper alternatives, *e.g.,* Ghost module [107], to supplant the regular residual blocks in our model for lightweight design. Our model also has the potential for all-in-one image restoration tasks due to the adaptive frequency learning ability for different degradation types.

REFERENCES

[1] Y. Lim, Y. Bliesener, S. Narayanan, and K. S. Nayak, "Deblurring for spiral real-time mri using convolutional neural networks," *Magnetic Resonance in Medicine*, vol. 84, no. 6, pp. 3438–3452, 2020.

[2] B. Rasti, Y. Chang, E. Dalsasso, L. Denis, and P. Ghamisi, "Image restoration for remote sensing: Overview and toolbox," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 201–230, 2021.

[3] S. Zang, M. Ding, D. Smith, P. Tyler, T. Rakotoarivelo, and M. A. Kaafar, "The impact of adverse weather conditions on autonomous vehicles: How rain, snow, fog, and hail affect the performance of a self-driving car," *IEEE Vehicular Technology Magazine*, vol. 14, no. 2, pp. 103–111, 2019.

[4] K. Zhang, W. Ren, W. Luo, W.-S. Lai, B. Stenger, M.-H. Yang, and H. Li, "Deep image deblurring: A survey," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2103–2130, 2022.

[5] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 4641–4650.

[6] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2034–2042.

[7] H. Son, J. Lee, S. Cho, and S. Lee, "Single image defocus deblurring using kernel-sharing parallel atrous convolutions," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 2642–2650.

[8] J. Li, W. Tan, and B. Yan, "Perceptual variousness motion deblurring with light global context refinement," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 4116–4125.

[9] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated fusion network for single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[10] H. Dong, J. Pan, L. Xiang, Z. Hu, X. Zhang, F. Wang, and M.-H. Yang, "Multi-scale boosted dehazing network with dense feature fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[11] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11 908–11 915.

[12] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 821–14 831.

[13] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 299–12 310.

[14] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 1833–1844.

[15] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 683–17 693.

[16] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[17] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *Proceedings of the European Conference on Computer Vision*, 2022.

[18] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua, "Gated context aggregation network for image dehazing and deraining," in *IEEE Winter Conference on Applications of Computer Vision*, 2019, pp. 1375–1383.

[19] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 304–16 313.

[20] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7314–7323.

[21] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[22] Y. Cui, Y. Tao, W. Ren, and A. Knoll, "Dual-domain attention for image deblurring," in *Association for the Advancement of Artificial Intelligence*, 2023.

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

[24] Y. Cui, Y. Tao, Z. Bing, W. Ren, X. Gao, X. Cao, K. Huang, and A. Knoll, "Selective frequency network for image restoration," in *International Conference on Learning Representations*, 2023.

[25] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[26] H. Gao, X. Tao, X. Shen, and J. Jia, "Dynamic scene deblurring with parameter selective sharing and nested skip connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[27] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[28] W.-T. Chen, H.-Y. Fang, C.-L. Hsieh, C.-C. Tsai, I. Chen, J.-J. Ding, S.-Y. Kuo *et al.*, "All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 4196–4205.

[29] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 111–126.

[30] K.-H. Liu, C.-H. Yeh, J.-W. Chung, and C.-Y. Chang, "A motion deblur method based on multi-scale high frequency residual image learning," *IEEE Access*, vol. 8, pp. 66 025–66 036, 2020.

[31] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, and Z. Jin, "Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 12 802–12 813.

[32] C. Wang, J. Pan, W. Wang, J. Dong, M. Wang, Y. Ju, J. Chen, and X.-M. Wu, "Promptrestorer: A prompting image restoration method with degradation perception," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[33] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2353–2363.

[34] Y. Cui, W. Ren, S. Yang, X. Cao, and A. Knoll, "Irnext: Rethinking convolutional network design for image restoration," in *Proceedings of the International Conference on Machine Learning*, 2023.

[35] J. Zhang, Y. Cao, Z.-J. Zha, and D. Tao, "Nighttime dehazing with a synthetic benchmark," in *Proceedings of the ACM International Conference on Multimedia*, 2020, pp. 2355–2363.

[36] W. Yan, R. T. Tan, and D. Dai, "Nighttime defogging using high-low frequency decomposition and grayscale-color networks," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 473–488.

[37] B. Huang, L. Zhi, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using sar image prior based on conditional generative adversarial networks," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1806–1813.

[38] Y. Liu, L. Zhu, S. Pei, H. Fu, J. Qin, Q. Zhang, L. Wan, and W. Feng, "From synthetic to real: Image dehazing collaborating with unlabeled real data," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 50–58.

[39] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: a dehazing benchmark with real hazy and haze-free outdoor images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 754–762.

[40] C. Ancuti, C. O. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-haze: A dehazing benchmark with real hazy and haze-free indoor images," in *Advanced Concepts for Intelligent Vision Systems: 19th International Conference, ACIVS 2018, Poitiers, France, September 24–27, 2018, Proceedings 19*, 2018, pp. 620–631.

[41] C. Li, H. Zhou, Y. Liu, C. Yang, Y. Xie, Z. Li, and L. Zhu, "Detection-friendly dehazing: Object detection in real-world hazy scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[42] S. Lee, T. Son, and S. Kwak, "Fifo: Learning fog-invariant features for foggy scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 911–18 921.

[43] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[44] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "Desnownet: Context-aware deep network for snow removal," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3064–3073, 2018.

[45] Y. Cui, W. Ren, and A. Knoll, "Omni-kernel network for image restoration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1426–1434.

[46] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Focal network for image restoration," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 13 001–13 011.

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

[48] H. Zhang, Y. Dai, H. Li, and P. Koniusz, "Deep stacked hierarchical multi-patch network for image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[49] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5812–5820.

[50] S. Chen, T. Ye, Y. Liu, T. Liao, J. Jiang, E. Chen, and P. Chen, "Msp-former: Multi-scale projection transformer for single image desnowing," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[51] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *arXiv preprint arXiv:2204.03883*, 2022.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[53] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 492–511.

[54] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, and J. Wang, "On the connection between local attention and dynamic depth-wise convolution," in *International Conference on Learning Representations*, 2021.

[55] S. Zhou, J. Zhang, J. Pan, H. Xie, W. Zuo, and J. Ren, "Spatio-temporal filter adaptive network for video deblurring," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[56] W. Wen, W. Ren, Y. Shi, Y. Nie, J. Zhang, and X. Cao, "Video super-resolution via a spatio-temporal alignment network," *IEEE Transactions on Image Processing*, vol. 31, pp. 1761–1773, 2022.

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

[58] B. K. Gunturk and X. Li, *Image restoration: fundamentals and advances*. CRC Press, 2012.

[59] M. R. Banham and A. K. Katsaggelos, "Digital image restoration," *IEEE Signal Processing Magazine*, vol. 14, no. 2, pp. 24–41, 1997.

[60] H. Yu, N. Zheng, M. Zhou, J. Huang, Z. Xiao, and F. Zhao, "Frequency and spatial dual guidance for image dehazing," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 181–198.

[61] M. Zhou, J. Huang, C.-L. Guo, and C. Li, "Fourmer: an efficient global modeling paradigm for image restoration," in *International Conference on Machine Learning*, 2023, pp. 42 589–42 601.

[62] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, 2005.

[63] N. Park and S. Kim, "How do vision transformers work?" in *International Conference on Learning Representations*, 2022.

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

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

[66] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[67] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[68] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma, "Contrastive learning for compact single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 551–10 560.

[69] T. Ye, Y. Zhang, M. Jiang, L. Chen, Y. Liu, S. Chen, and E. Chen, "Perceiving and modeling density for image dehazing," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 130–145.

[70] B. Li, X. Peng, Z. Wang, J. Xu, and D. Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[71] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Image restoration via frequency selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[72] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images," in *IEEE International Conference on Image Processing*, 2019, pp. 1014–1018.

[73] C. O. Ancuti, C. Ancuti, and R. Timofte, "Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[74] X. Chen, Y. Li, L. Dai, and C. Kong, "Hybrid high-resolution learning for single remote sensing satellite image dehazing," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[75] Y. Li and X. Chen, "A coarse-to-fine two-stage attentive network for haze removal of remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 10, pp. 1751–1755, 2020.

[76] Y. Zheng, J. Zhan, S. He, J. Dong, and Y. Du, "Curricular contrastive regularization for physics-aware single image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5785–5794.

[77] K. Chi, Y. Yuan, and Q. Wang, "Trinity-net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[78] A. Kulkarni, S. S. Phutke, and S. Murala, "Unified transformer network for multi-weather image restoration," in *European Conference on Computer Vision*, 2022, pp. 344–360.

[79] Y. Jin, B. Lin, W. Yan, Y. Yuan, W. Ye, and R. T. Tan, "Enhancing visibility in nighttime haze images using guided apsf and gradient adaptive convolution," in *Proceedings of the ACM International Conference on Multimedia*, 2023, pp. 2446–2457.

[80] J. Zhang, Y. Cao, and Z. Wang, "Nighttime haze removal based on a new imaging model," in *IEEE International Conference on Image Processing*, 2014, pp. 4557–4561.

[81] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 226–234.

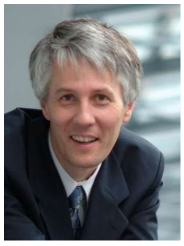
[82] J. Zhang, Y. Cao, S. Fang, Y. Kang, and C. Wen Chen, "Fast haze removal for nighttime image using maximum reflectance prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[83] T. Wang, G. Tao, W. Lu, K. Zhang, W. Luo, X. Zhang, and T. Lu, "Restoring vision in hazy weather with hierarchical contrastive learning," *Pattern Recognition*, vol. 145, p. 109956, 2024.
[84] C. Ancuti, C. O. Ancuti, C. De Vleeschouwer, and A. C. Bovik, "Night-time dehazing by fusion," in *IEEE International Conference on Image Processing*, 2016, pp. 2256–2260.
[85] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.
[86] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 754–770.
[87] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "Desnownet: Context-aware deep network for snow removal," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3064–3073, 2018.
[88] R. Li, R. T. Tan, and L.-F. Cheong, "All in one bad weather removal using architectural search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
[89] B. Cheng, J. Li, Y. Chen, S. Zhang, and T. Zeng, "Snow mask guided adaptive residual network for image snow removal," *arXiv preprint arXiv:2207.04754*, 2022.
[90] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3943–3956, 2019.
[91] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[92] X. Fu, J. Huang, X. Ding, Y. Liao, and J. Paisley, "Clearing the skies: A deep network architecture for single-image rain removal," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2944–2956, 2017.
[93] W. Wei, D. Meng, Q. Zhao, Z. Xu, and Y. Wu, "Semi-supervised transfer learning for image rain removal," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[94] R. Yasarla and V. M. Patel, "Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[95] X. Li, J. Wu, Z. Lin, H. Liu, and H. Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *Proceedings of the European Conference on Computer Vision*, 2018.
[96] D. Ren, W. Zuo, Q. Hu, P. Zhu, and D. Meng, "Progressive image deraining networks: A better and simpler baseline," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
[97] K. Jiang, Z. Wang, P. Yi, C. Chen, B. Huang, Y. Luo, J. Ma, and J. Jiang, "Multi-scale progressive fusion network for single image deraining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
[98] A. Abuolaim, M. Afifi, and M. S. Brown, "Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2022, pp. 1231–1239.
[99] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[100] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
[101] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown, "Rain streak removal using layer priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
[102] J. Rim, G. Kim, J. Kim, J. Lee, S. Lee, and S. Cho, "Realistic blur synthesis for learning image deblurring," in *Proceedings of the European Conference on Computer Vision*, 2022, pp. 487–503.
[103] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[104] J. Zhou, P. Wang, F. Wang, Q. Liu, H. Li, and R. Jin, "Elsa: Enhanced local self-attention for vision transformer," *arXiv preprint arXiv:2112.12786*, 2021.
[105] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
[106] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 764–773.
[107] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1580–1589.

Yuning Cui (Student Member, IEEE) received the B.Eng. degree from Central South University, China, in 2016, and the M.Eng. degree from National University of Defense Technology, China, in 2018. He is currently working towards the Ph.D. degree at the Chair of Robotics, Artificial Intelligence and Real-time Systems within the School of Computation, Information and Technology at the Technical University of Munich. His research interest lies in image restoration.

Wenqi Ren (Member, IEEE) received the Ph.D. degree from Tianjin University, Tianjin, China, in 2017. From 2015 to 2016, he was supported by the China Scholarship Council and working with Prof. Ming-Husan Yang as a Joint-Training Ph.D. Student with the Electrical Engineering and Computer Science Department, University of California at Merced. He is currently a Professor with the School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China. His research interests include image processing and related high-level vision problems. He received the Tencent Rhino Bird Elite Graduate Program Scholarship in 2017 and the MSRA Star Track Program in 2018.

Xiaochun Cao (Senior Member, IEEE) received the B.E. and M.E. degrees in computer science from Beihang University, Beijing, China, and the Ph.D. degree in computer science from the University of Central Florida, Orlando, FL, USA. He has been a Professor with School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen, China. After graduation, he spent about three years at OjectVideo Inc., as a Research Scientist. From 2008 to 2012, he was a Professor with Tianjin University, Tianjin, China. He is a fellow of IET. He is on the Editorial Board of the IEEE Transactions on Image Processing. His dissertation was nominated for the University of Central Florida's University-Level Outstanding Dissertation Award.



Alois Knoll (Fellow, IEEE) received his diploma (M.Sc.) degree in Electrical/Communications Engineering from the University of Stuttgart, Germany, in 1985 and his Ph.D. (*summa cum laude*) in Computer Science from Technical University of Berlin, Germany, in 1988. He served on the faculty of the Computer Science department at TU Berlin until 1993. He joined the University of Bielefeld, Germany as a full professor and served as the director of the Technical Informatics research group until 2001. Since 2001, he has been a professor at the Department of Informatics, Technical University of Munich (TUM), Germany. His research interests include cognitive, medical and sensor-based robotics, multi-agent systems, data fusion, adaptive systems, multimedia information retrieval, model-driven development of embedded systems with applications to automotive software and electric transportation, as well as simulation systems for robotics and traffic.