IRNeXt: Rethinking Convolutional Network Design for Image Restoration

Yuning Cui [1] **Wenqi Ren** [2 3] **Sining Yang** [3] **Xiaochun Cao** [3] **Alois Knoll** [1]

Abstract

We present IRNeXt, a simple yet effective convolutional network architecture for image restoration. Recently, Transformer models have dominated the field of image restoration due to the powerful ability of modeling long-range pixels interactions. In this paper, we excavate the potential of the convolutional neural network (CNN) and show that our CNN-based model can receive comparable or better performance than Transformer models with low computation overhead on several image restoration tasks. By re-examining the characteristics possessed by 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 IRNeXt delivers state-of-the-art performance among numerous datasets on a range of image restoration tasks with low computational complexity, including image dehazing, single-image defocus/motion deblurring, image deraining, and image desnowing. https://github.com/c-yn/IRNeXt.

1. Introduction

Image restoration aims to restore a clean image from its degraded counterpart, playing an essential role in remote sensing, self-driving techniques, photography, and medical imaging (Lim et al., 2020; Rasti et al., 2021; Zang et al., 2019). 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 (Zhang et al., 2022).

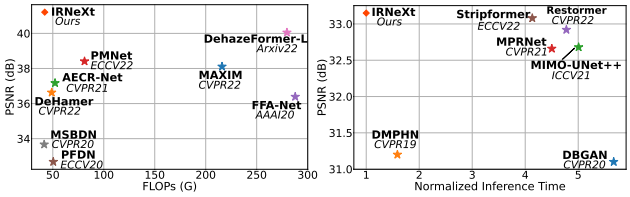

Figure 1. Comparisons between our IRNeXt and other state-of-the-art algorithms. **Left**: PSNR *vs.* FLOPs on SOTS-Indoor (Li et al., 2018a) for dehazing. **Right**: PSNR *vs.* normalized inference time (by ours) on GoPro (Nah et al., 2017) for motion deblurring.

With the rapid development of deep learning, multifarious CNN-based methods have been proposed based on ingenious modules or borrowed units, such as encoder-decoder architecture (Cho et al., 2021; Lee et al., 2021), dilated convolution (Son et al., 2021; Li et al., 2021; Ren et al., 2018), and attention mechanisms (Qin et al., 2020; Zamir et al., 2021). Recent years have witnessed a paradigm shift from CNN-based architectures to Transformer models (Chen et al., 2021a; Liang et al., 2021). These models have significantly advanced the performance of image restoration. However, how to reduce the complexity of self-attention for image restoration is still a non-trivial problem.

Our main goal is to exploit an efficient and effective image restoration architecture based on CNNs. By delving into previous advanced image restoration methods, we summarize several critical factors that a successful image restoration model should have as follow: **(a)** Multi-scale representation learning. Recent deep architectures resort to a single encoder-decoder (Chen et al., 2019; Lee et al., 2021; Ruan et al., 2022) or multi-stage paradigm (Zamir et al., 2021; Ren et al., 2018; Liu et al., 2019) to learn multi-scale feature representations, which are helpful for removal of 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 (Qin et al., 2020; Suin et al., 2020; Cui et al., 2023b). **(c)** Frequency modulation. Frequency modulation operation is a powerful complement to the spatial feature refinement by reducing the frequency discrepancy between sharp and degraded images (Zou et al., 2021; Cui et al., 2023a). **(d)** Low computational complexity. This is essential for image restoration, which often involves high-resolution images.

[1]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany [2]School of Ocean Information Engineering, Jimei Univeristy, Xiamen, China [3]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, China. Correspondence to: Wenqi Ren <renwq3@mail.sysu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

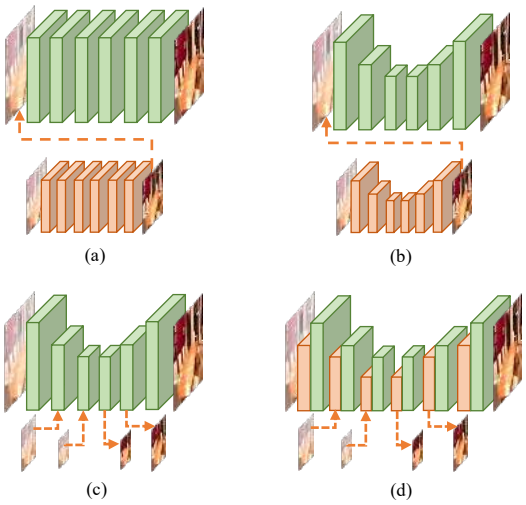

Figure 2. Comparisons between multi-scale architectures. **(a)** Multi-stage isotropic design (Ren et al., 2018; Nah et al., 2017). **(b)** Multi-stage U-shaped design (Gao et al., 2019). **(c)** Single U-shaped design (Chen et al., 2019; Lee et al., 2021). **(d)** Our network that imitates the multi-stage design in a U-shaped pipeline.

Taking into account 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 2, and propose imitating the multi-stage mechanism in a single U-shaped network. Specifically, for each scale, we downsample the feature map into different sizes such that the model is capable of handling blurs in a coarse-to-fine manner. **Secondly**, inspired by (Han et al., 2021), we develop a local attention module, which not only can perform information aggregation efficiently, but is adaptive to the input feature. **Thirdly**, since frequency discrepancies between sharp and degraded image pairs mostly lie in the high-frequency components (Liu et al., 2020), we accentuate the informative frequency part for restoration by recalibrating the weight of the high-pass filter in the obtained attention map of the local attention module. **Finally**, we insert above modules into a convolutional U-shaped backbone to establish IRNeXt. Our contributions can be summarized as follow:

- We identify the properties that a successful image restoration method possesses and propose a novel convolutional model, dubbed IRNeXt, which enhances multi-scale representation learning by incorporating the multi-stage mechanism into a U-shaped network.
- We present an efficient content-aware local attention module that can emphasize the useful frequency bands by reweighing the weight of the high-pass filter.
- Extensive experiments demonstrate that our model delivers state-of-the-art performance on 13 benchmark datasets for five typical image restoration tasks.

2. Related Work

Image Restoration. As a long-standing task, image restoration provides high-quality images for visibility and downstream high-level tasks. CNNs have become the mainstream in this field for several years and achieved many successful stories on various restoration tasks (Nah et al., 2017; Cai et al., 2016; Liu et al., 2018b; Abuolaim & Brown, 2020). To boost the performance, many advanced modules have been developed to strengthen the ability of these CNN-based frameworks, such as encoder-decoder architecture (Cho et al., 2021; Lee et al., 2021), multi-stage paradigm (Nah et al., 2017; Gao et al., 2019; Tu et al., 2022), multi-patch learning (Zhang et al., 2019a; Suin et al., 2020; Zamir et al., 2021), and attention mechanisms (Qin et al., 2020; Anwar & Barnes, 2019; Zhang et al., 2018). Recently, numerous Transformer models have been proposed to capture long-range dependencies, and have significantly advanced the state-of-the-art performance of image restoration (Chen et al., 2021a; Liang et al., 2021; Guo et al., 2022; Song et al., 2022). Despite a few remedies (Zamir et al., 2022; Wang et al., 2022; Tsai et al., 2022), however, how to reduce the complexity of self-attention remains formidable.

Attention Mechanisms. Driven by the success of attention mechanisms in high-level tasks, various attention modules have been proposed to attend to important contents for image restoration (Zamir et al., 2020; Qin et al., 2020; Suin et al., 2020; Liu et al., 2019). Our local attention module mimics the depth-wise convolution (Han et al., 2021) to conduct information aggregation, which has the content-aware property as self-attention while remaining computationally efficient. The most related works to our module are the methods that learn the dynamic filter for restoration (Lee et al., 2021; Zhou et al., 2019; Wen et al., 2022). Different from this kind of approaches, the proposed module does not produce as many attention weights as them, leading to fewer parameters and lower complexity. Furthermore, instead of directly imposing attention weights on the input feature, we perform filter modulation in advance to accentuate the informative spectral part of feature by rescaling the weight of the high-pass filter in the attention map.

Spectral Networks. There is a big difference between the spectral features of clean/degraded image pairs (Liu et al., 2020; Mao et al., 2021). A few deep restoration frameworks have taken measures to bridge this gap by refining features in the frequency domain. The common practice is first to transform spatial features into the spectral domain via wavelet or Fourier transform, and then leverage convolutions to modulate features (Zou et al., 2021; Yu et al., 2022; Chen et al., 2021c). Different from previous methods, IRNeXt performs filter modulation on the attention weights. Furthermore, our method does not include extra convolution layers and inverse transform, *e.g.,* inverse Fourier transform.

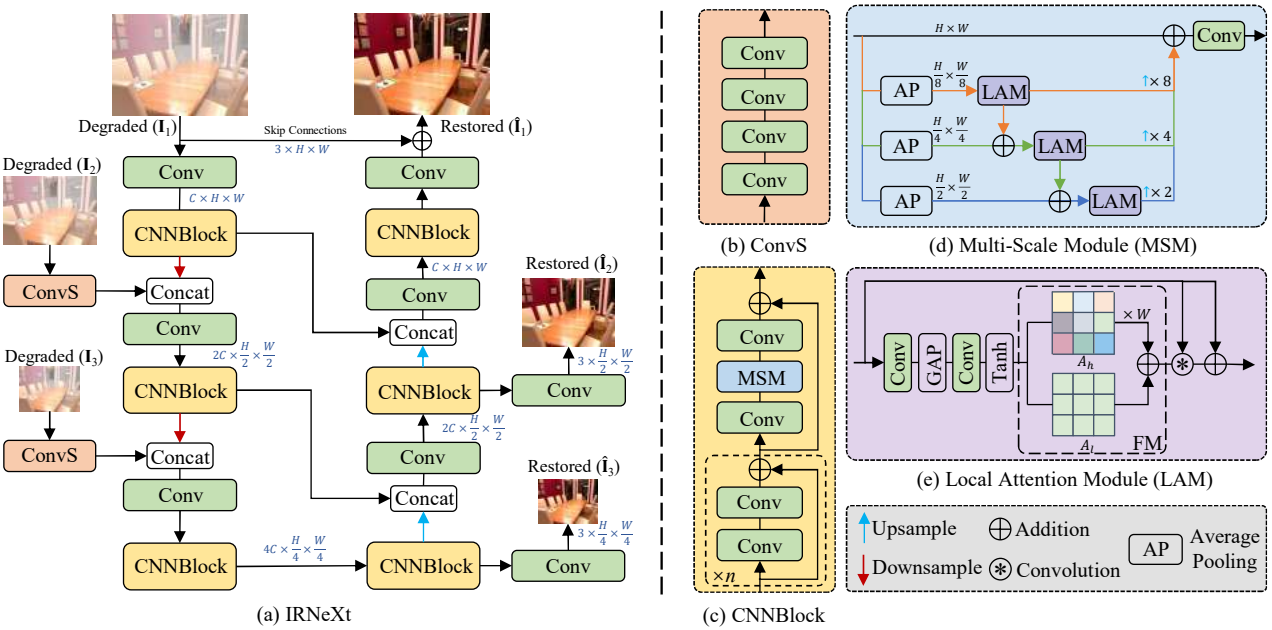

Figure 3. The architecture of the proposed IRNeXt. **(a)** IRNeXt consists of six CNNBlocks and adopts the multi-input and multi-output strategies for image restoration. **(b)** ConvS extracts the shallow features for low-resolution degraded images, which includes a series of convolutions with kernel sizes of 3×3, 1×1, 3×3, and 1×1. **(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. **(e)** Local attention module (LAM) performs information aggregation based on filter modulation (FM).

3. Method

In this section, we first describe the overall pipeline of IRNeXt. Then we present the core components of IRNeXt: multi-scale module (MSM) and local attention module (LAM). The loss functions are introduced in the final part.

3.1. Overall Architecture

As illustrated in Figure 3 (a), the proposed IRNeXt adopts a single U-shaped architecture for image restoration. Specifically, given any degraded image $\mathbf{I} \in \mathbb{R}^{3 \times H \times W}$, IRNeXt 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 obtain the high-level 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 (Cho et al., 2021; Tu et al., 2022; Mao et al., 2021), multiple downsampled degraded images are merged into the main path via the ConvS (Figure 3 (b)) module and concatenation, followed by a 3×3 convolution to adjust the number of channels. Next, the high-level features are fed into another three CNNBlocks to restore the high-resolution features. The multi-output strategy is adopted, where the low-resolution predicted image is generated using a 3×3 convolution and image-level skip connection (omitted in Figure 3 (a) for simplicity) after the first two CNNBlocks. Furthermore, decoder features are concatenated with the encoder features to assist restoration, and a 1×1 convolution layer is used to reduce the number of channels by half. The final clean image is produced after adding the original degraded input. Next, we detail the proposed modules: MSM and LAM.

3.2. Multi-Scale Module

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 are limited to handle degradation blurs of different sizes. To enhance multi-scale learning and remove blurs in a coarse-to-fine manner at each scale, we mimic the multi-stage network and implement it in each scale of a single U-shaped framework, as illustrated in Figure 2 (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) operations with different downsampling ratios to convert $\mathbf{X}$ into distinct scale spaces. For each branch, the resulting feature after LAM is incorporated into the next branch via addition operation. In this way, MSM is capable of removing degradations in a progressive manner by imitating

the multi-stage network. Finally, the output features of all branches are unified to the input size and added together. In IRNeXt, 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 connection), the output features can be obtained by:

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

where $\hat{\mathbf{X}}_0$=**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:

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

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

3.3. Local Attention Module

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 (Chen et al., 2021a; Liang et al., 2021). Despite a few remedies (Tsai et al., 2022; Wang et al., 2022), 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 (Zamir et al., 2022).

In this work, we present LAM by combining the merits of both the self-attention and convolution operator. Our LAM inherits the content-aware property of the former, and maintains the efficiency characteristic of the latter. More concretely, LAM leverages a simple convolution block to generate attention weights, which are adaptive to the input feature, and then performs aggregation using 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 (Park & Kim, 2022), which is unsuitable for image restoration, because the large discrepancies between the sharp and degraded images mostly lie in the high-frequency components (Liu et al., 2020; Mao et al., 2021).

In LAM, we resolve the above issue in the attention weights generation step from two aspects: (*i*) breaking through the limitation of the low-pass filter using Tanh; and (*ii*) emphasizing the weight of high-pass filters in attention weights using the proposed filter modulation (FM) operation.

Utilization of Tanh. We simply substitute the Tanh function for Softmax. This scheme enjoys two advantages. Firstly, we break through 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, the attention weights generating process can be expressed as:

$$\mathbf{A} = \mathrm{Tanh}(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(\mathrm{Conv}_{3\times3}(\mathbf{X})))) \quad (3)$$

where GAP denotes the global average pooling and Tanh represents the hyperbolic tangent function. To strike a better tradeoff between the complexity and diversity of attention weights, instead of producing attention weights for each channel (Zhou et al., 2019; Lee et al., 2021), we impose attention weights on the input feature in groups. In each feature group, attention weights are shared across both channel and spatial dimensions. $\mathbf{A} \in \mathbb{R}^{G\times K\times K}$, where G is the number of groups and K^2 is the region size for integration.

FM. In addition to the utilization of Tanh, we further propose lifting the weight of high-pass filters in the attention map. To this end, as illustrated in Figure 3 (e), we first decompose the attention map $\mathbf{A}$ 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 $\mathbf{A}$ by:

$$\mathbf{A}_l = \frac{1}{K^2}\mathbf{E} \quad (4)$$

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

$$\mathbf{A}_h = \mathbf{A} - \mathbf{A}_l \quad (5)$$

Next, the modulated attention map can be obtained by:

$$\mathbf{A}' = \mathbf{A}_l + W\mathbf{A}_h \quad (6)$$

where W denotes the learnable parameters, which are directly optimized by backpropagation and initialized as **1**. It is worth mentioning that our design is extremely lightweight, as it does not introduce additional convolution layers as other frequency-based networks (Zou et al., 2021; Mao et al., 2021).

Finally, we apply the resulting attention weights to the input feature via the convolution operation. For features in each group, the output can be obtained by:

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

where g is the index of the group.

3.4. Loss Functions

Since we introduce FM in LAM, apart from the spatial $\mathcal{L}_1$ loss, we adopt the spectral $\mathcal{L}_1$ loss to accentuate the useful

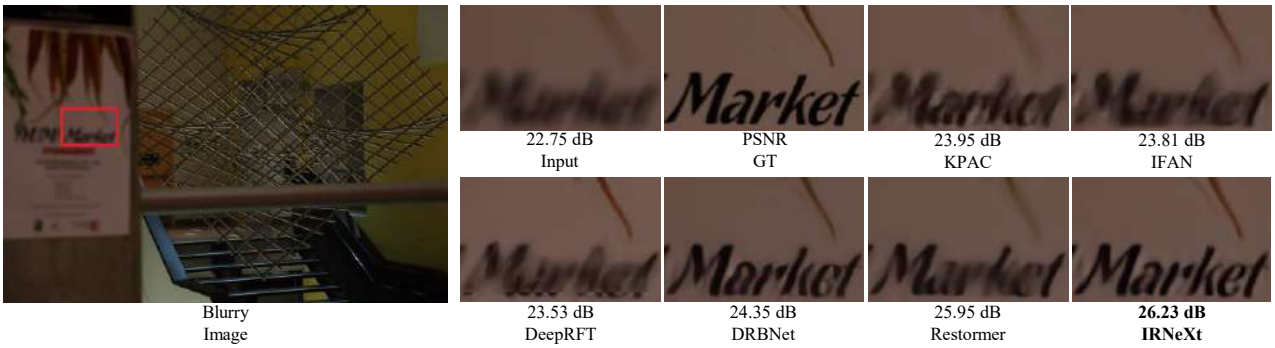

Figure 4. Single-image defocus deblurring comparisons on the DPDD (Abuolaim & Brown, 2020) dataset.

Method	Indoor Scenes				Outdoor Scenes				Combined			
	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓
DPDNet (Abuolaim & 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**
IRNeXt (Ours)	**29.22**	0.879	**0.024**	0.167	**23.53**	**0.752**	**0.049**	0.244	**26.30**	**0.814**	**0.037**	0.206

Table 1. Single-image defocus deblurring comparisons on the DPDD (Abuolaim & Brown, 2020) dataset.

frequency. The dual-domain loss functions are given by:

$$\mathcal{L}_{spatial} = \sum_{i=1}^{3} \frac{1}{P_i} \|\hat{\mathbf{I}}_i - \mathbf{Y}_i\|_1 \tag{8}$$

$$\mathcal{L}_{frequency} = \sum_{i=1}^{3} \frac{1}{P_i} \|F(\hat{\mathbf{I}}_i) - F(\mathbf{Y}_i)\|_1 \tag{9}$$

where i is the index of multiple outputs as shown in Figure 3 (a); $\hat{\mathbf{I}}$ and $\mathbf{Y}$ represent the predicted image and ground truth; P denotes the total elements of the image for normalization; and F is the fast Fourier transform. The final loss function is given by integrating above two terms:

$$\mathcal{L}_{total} = \mathcal{L}_{spatial} + \lambda \mathcal{L}_{frequency} \tag{10}$$

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

4. Experiments

To demonstrate the effectiveness of our model, we evaluate IRNeXt on 13 datasets for five image restoration tasks: single-image defocus deblurring, image dehazing, image deraining, image desnowing, and image motion deblurring. We provide more details of the used datasets, training settings, and additional visual results in Appendix. In tables, the best result is highlighted in **bold**.

Implementation details. We train separate models for different problems. In all experiments, unless specified otherwise, the following hyper-parameters are used. We choose $G = 8$ and $K = 3$ in Eq. 7. We train our model using the Adam optimizer (Kingma & Ba, 2014) with the initial learning rate as $1e^{-4}$, which is gradually reduced to $1e^{-6}$ with cosine annealing (Loshchilov & Hutter, 2016). For data augmentation, we use random horizontal flips. With the exception of GoPro (Nah et al., 2017), where n in Figure 3 (c) is set as 13, we set $n = 15$ for deraining and deblurring tasks, and $n = 3$ for dehazing and desnowing datasets. All models are trained and evaluated on an NVIDIA Tesla V100 GPU. FLOPs are measured on 256×256 patches.

4.1. Single-Image Defocus Deblurring Results

We conduct single-image defocus deblurring experiments on the widely used DPDD (Abuolaim & Brown, 2020) dataset. The image fidelity scores are presented in Table 1. Our model obtains the highest scores on most categories. Concretely, IRNeXt outperforms Restormer (Zamir et al., 2022) by ∼0.3 dB PSNR in all scenes. Compared with the frequency-based DeepRFT (Mao et al., 2021), IRNeXt shows a substantial gain of 0.59 dB PSNR on the combined scene category. Figure 4 illustrates that IRNeXt recovers the sharper and visually-faithful image than other approaches.

4.2. Image Dehazing Results

We evaluate IRNeXt on the synthetic dataset (RESIDE/SOTS (Li et al., 2018a)) and real-world datasets (NH-HAZE (Ancuti et al., 2020) and Dense-Haze (Ancuti et al.,

Figure 5. Image dehazing comparisons on the SOTS-Indoor (Li et al., 2018a) dataset.

Figure 6. Image deraining comparisons on the Rain100H (Yang et al., 2017) dataset.

Method	SOTS-Indoor		SOTS-Outdoor		Dense-Haze		NH-HAZE		#Param (M)
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
AOD-Net (Li et al., 2017)	20.51	0.816	24.14	0.920	13.14	0.414	15.40	0.569	0.002
GridDehazeNet (Liu et al., 2019)	32.16	0.984	30.86	0.982	13.31	0.368	13.80	0.537	0.956
MSBDN (Dong et al., 2020)	33.67	0.985	33.48	0.982	15.37	0.486	19.23	0.706	31.35
PFDN (Dong & Pan, 2020)	32.68	0.976	-		-		-		11.27
FFA-Net (Qin et al., 2020)	36.39	0.989	33.57	0.984	14.39	0.452	19.87	0.692	4.456
KDDN (Hong et al., 2020)	34.72	0.985	33.57	0.984	14.28	0.407	17.39	0.590	5.99
AECR-Net (Wu et al., 2021)	37.17	0.990	-		15.80	0.466	19.88	0.717	2.611
DeHamer (Guo et al., 2022)	36.63	0.988	35.18	0.986	16.62	0.560	**20.66**	0.684	132.45
DehazeFormer-L (Song et al., 2022)	40.05	**0.996**	-		-		-		25.44
MAXIM (Tu et al., 2022)	38.11	0.991	34.19	0.985	-		-		14.1
FSDGN (Yu et al., 2022)	38.63	0.990	-		16.91	0.581	19.99	0.731	2.73
PMNet (Ye et al., 2022)	38.41	0.990	34.74	0.985	16.79	0.510	20.42	0.730	18.90
IRNeXt (Ours)	**41.21**	**0.996**	**39.18**	**0.996**	**17.60**	**0.659**	20.55	**0.813**	5.46

Table 2. Image dehazing comparisons on the synthetic dataset (RESIDE/SOTS (Li et al., 2018b)) and real-world datasets (Dense-Haze (Ancuti et al., 2019) and NH-HAZE (Ancuti et al., 2020)).

Method	Rain100L		Rain100H		Test100	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DerainNet (Fu et al., 2017a)	27.03	0.884	14.92	0.592	22.77	0.810
DIDMDN (Wei et al., 2019)	25.23	0.741	17.35	0.524	22.56	0.818
UMRL (Yasarla & Patel, 2019)	29.18	0.923	26.01	0.832	24.41	0.829
RESCAN (Li et al., 2018b)	29.80	0.881	26.36	0.786	25.00	0.835
PreNet (Ren et al., 2019)	32.44	0.950	26.77	0.858	24.81	0.851
MSPFN (Jiang et al., 2020)	32.40	0.933	28.66	0.860	27.50	0.876
MPRNet (Zamir et al., 2021)	36.40	0.965	30.41	0.890	30.27	0.897
HINet (Chen et al., 2021b)	37.20	0.969	30.63	0.893	30.26	0.905
DRT (Liang et al., 2022)	37.61	0.948	29.47	0.846	27.02	0.847
MAXIM (Tu et al., 2022)	38.06	**0.977**	30.81	**0.903**	31.17	**0.922**
IRNeXt (Ours)	**38.14**	0.972	**31.64**	0.902	**31.53**	0.919

Table 3. Image deraining quantitative comparisons on three widely used datasets: Rain100L (Yang et al., 2017), Rain100H (Yang et al., 2017), and Test100 (Zhang et al., 2019b).

2019)). The results are shown in Table 2. Our model receives the best performance on most metrics. Particularly on the SOTS-Outdoor scene category, IRNeXt significantly outperforms the Transformer model DeHamer (Guo et al., 2022) by 4 dB in terms of PSNR, while having 24× fewer parameters. On the SOTS-Indoor category, our method receives a performance gain of 1.16 dB PSNR over the recent algorithm DehazeFormer (Song et al., 2022) with 85% fewer FLOPs, as illustrated in Figure 1 (Left). Figure 5 depicts that our result is visually closer to the ground truth.

We further compare IRNeXt with other state-of-the-art approaches on the two real-world datasets. IRNeXt is more effective in removing real hazy blurs than other methods, outperforming the recent algorithm (Yu et al., 2022) by 0.81 and 0.13 dB on Dense-Haze and NH-HAZE, respectively.

4.3. Image Deraining Results

Following previous methods (Zamir et al., 2021; Tu et al., 2022), we train IRNeXt on a composite dataset and compute the metrics on the Y channel in YCbCr color space. Table 3 shows the results on Rain100L (Yang et al., 2017),

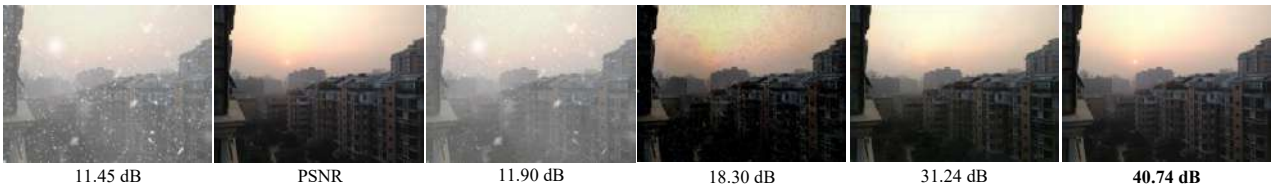

Figure 7. Image desnowing comparisons on the CSD (Chen et al., 2021c) dataset.

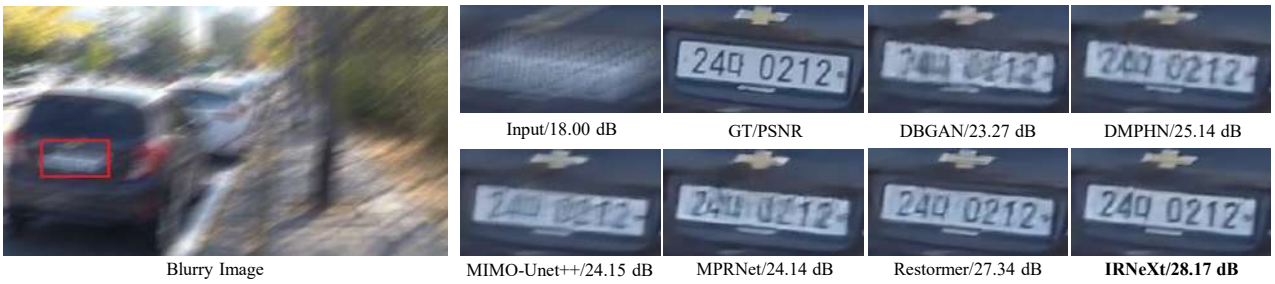

Figure 8. Image motion deblurring comparisons on the GoPro (Nah et al., 2017) dataset.

Method	CSD		SRRS		Snow100K	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DesnowNet (Liu et al., 2018a)	20.13	0.81	20.38	0.84	30.50	0.94
CycleGAN (Engin et al., 2018)	20.98	0.80	20.21	0.74	26.81	0.89
All in One (Li et al., 2020)	26.31	0.87	24.98	0.88	26.07	0.88
JSTASR (Chen et al., 2020)	27.96	0.88	25.82	0.89	23.12	0.86
HDCW-Net (Chen et al., 2021c)	29.06	0.91	27.78	0.92	31.54	0.95
SMGARN (Cheng et al., 2022)	31.93	0.95	29.14	0.94	31.92	0.93
MSP-Former (Chen et al., 2022)	33.75	0.96	30.76	0.95	33.43	**0.96**
TransWeather (Valanarasu et al., 2022)	31.76	0.93	28.29	0.92	31.82	0.93
IRNeXt (Ours)	**37.29**	**0.99**	**31.91**	**0.98**	**33.61**	0.95

Table 4. Image desnowing comparisons on the CSD (Chen et al., 2021c), SRRS (Chen et al., 2020), and Snow100K (Liu et al., 2018a) datasets.

Method	PSNR	SSIM
SRN-DeblurNet (Tao et al., 2018)	32.53	0.840
MIMO-UNet (Cho et al., 2021)	32.73	0.846
MIMO-UNet+ (Cho et al., 2021)	33.37	0.856
MPRNet (Zamir et al., 2021)	33.61	0.861
Restormer (Zamir et al., 2022)	33.69	0.863
Uformer (Wang et al., 2022)	33.98	0.866
IRNeXt (Ours)	**34.08**	**0.869**

Table 5. Image motion deblurring numerical comparisons on the lately proposed real-world dataset RSBlur (Rim et al., 2022). The proposed model IRNeXt advances the state-of-the-art performance by 0.1 dB in terms of PSNR.

Method	PSNR	SSIM	FLOPs/G	Params/M	Time/s
DBGAN (Zhang et al., 2020)	31.10	0.942	759.85	**11.6**	1.447
DMPHN (Zhang et al., 2019a)	31.20	0.940	-	21.7	0.405
MIMO-UNet++ (Cho et al., 2021)	32.68	0.959	617.64	16.1	1.277
MPRNet (Zamir et al., 2021)	32.66	0.959	777.01	20.1	1.148
Restormer (Zamir et al., 2022)	32.92	0.961	140.99	26.1	1.218
Stripformer (Tsai et al., 2022)	33.08	**0.962**	170.46	20.0	1.054
IRNeXt (Ours)	**33.16**	**0.962**	**114.79**	13.21	**0.255**

Table 6. Image motion deblurring results on the GoPro (Nah et al., 2017) dataset. The inference time is tested in a synchronized manner by using *torch.cuda.synchronize*.

Rain100H (Yang et al., 2017), and Test100 (Zhang et al., 2019b). Compared with other algorithms, our method receives the comparable or better performance. On Rain100H, IRNeXt obtains a performance gain of 0.83 dB over MAXIM (Tu et al., 2022). Compared with the multi-stage method MPRNet (Zamir et al., 2021), our method yields a 1.41 dB improvement when averaged across all datasets. Figure 6 illustrates that IRNeXt produces a visually pleasant image on Rain100H.

4.4. Image Desnowing Results

We compare desnowing performance on three datasets: CSD (Chen et al., 2021c), SRRS (Chen et al., 2020), and Snow100K (Liu et al., 2018a). The numerical results are reported in Table 4. Our method significantly outperforms other algorithms on most categories. In particular, on the lately proposed CSD dataset, IRNeXt yields a 5.53 dB improvement over TransWeather (Valanarasu et al., 2022). Furthermore, compared with MSP-Former (Chen et al., 2022), which is elaborately designed for desnowing, our model shows a significant performance boost of 3.54 dB on the CSD dataset. Figure 7 illustrates that IRNeXt produces snow-free image without artifacts.

4.5. Image Motion Deblurring Results

We evaluate our model on the synthetic dataset GoPro (Nah et al., 2017) and real-world dataset RSBlur (Rim et al., 2022). The overall comparisons in terms of accuracy and

Method	PSNR	Params/M	FLOPs/G	Time/s
(a)Baseline	31.23	6.90	66.32	0.134
(b)Baseline+MSM+Conv	31.46	8.45	71.17	0.152
(c)Baseline+MSM+LAM w/o FM	31.53	8.55	71.19	0.165
(d)Full	31.64	8.56	71.19	0.166

Table 7. Break-down ablation study toward better performance. To separately study the effect of MSM, we deploy a 3×3 convolution in each branch to form variant (b). Here, MSM denotes the pure multi-scale paradigm without LAM.

8	4	2	PSNR	Params/M	FLOPs/G
			31.23	6.90	66.32
		✓	31.33	7.71	70.89
	✓	✓	31.39	8.13	71.12
✓	✓		31.51	8.13	70.26
✓	✓	✓	31.64	8.56	71.19

Table 8. The number of branches in MSM. The number indicates the downsampling rate of a branch.

Pooling Type	PSNR	Params
Convolution	31.51	8.58
Max Pooling	31.46	8.56
Average Pooling	31.64	8.56

Table 9. Ablation study on different pooling operations.

Method	PSNR
Softmax	31.50
Sigmoid	31.58
Tanh	31.64

Table 10. Ablation on different functions.

computational costs on GoPro are shown in Table 6. Compared with Transformer models Restormer (Zamir et al., 2022) and Stripformer (Tsai et al., 2022), IRNeXt receives performance gains of 0.24 dB and 0.08 dB respectively with fewer parameters and lower complexity. Furthermore, our model runs 4.78× and 4.13× faster than these two algorithms, respectively, demonstrating that our model strikes a better tradeoff between the accuracy and computation burden. In addition, we report the results on RSBlur in Table 5. As can be seen, our method outperforms the previous best approach Uformer (Wang et al., 2022) by 0.1 dB PSNR and 0.003 SSIM. Figure 8 illustrates that IRNeXt reconstructs more details than others for the difficult example of GoPro.

5. Ablation Study

Following (Tu et al., 2022), we conduct ablation studies on the GoPro (Nah et al., 2017) dataset with $n = 7$ (Figure 3 (c)). The baseline is obtained by removing MSM from our network. All models are trained for 1000 epochs. In tables, the final choices of IRNeXt are highlighted in gray.

Break-down ablation. We perform the break-down ablation by applying the proposed modules to the baseline successively. The results are reported in Table 7. The baseline model receives 31.23 dB (Table 7a). After deploying MSM with a 3×3 convolution in each branch, the model achieves a 0.23 dB improvement (Table 7b). Substituting LAM without FM for the convolution, the model obtains a further boosted performance of 0.07 dB (Table 7c). When using all components, the model achieves the best performance, 0.41 dB higher than that of the baseline, and only introduces 1.66 M parameters and 4.87 G FLOPs (Table 7d). The results demonstrate the effectiveness of our modules.

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 in MSM. The results are presented in Table 8. Employing more branches leads to better performance. Specifically, when using a single branch with the downsampling rate as 2, the model receives a gain of 0.1 dB over the baseline. When equipped with three branches, the model produces a 0.41 dB improvement, which is 0.03 dB higher than the sum of performance gains brought by separately using the branches 8,4 and branch 2, demonstrating the efficacy of the proposed 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 9. The average pooling variant achieves 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 apply Tanh in LAM to generate the negative weights for detrimental pixels when performing information aggregation. The comparisons are reported in Table 10. Compared with Softmax, Sigmoid receives a gain of 0.08 dB by breaking through the sum-to-one property. Tanh projects attention weights into (-1, 1), producing a further improvement of 0.06 dB. Hence, we choose to use Tanh in the final model.

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 a simple convolutional model IRNeXt for image restoration. Specifically, we introduce a multi-scale module (MSM), which incorporates the multi-stage mechanism into a single U-shaped network to remove blurs of different sizes in a coarse-to-fine manner. Moreover, we propose an efficient local attention module (LAM) for handling spatially-varying blurs, where informative spectral features are accentuated via filter modulation (FM). Extensive experiments on 13 benchmark datasets demonstrate that the proposed IRNeXt achieves state-of-the-art performance for several image restoration tasks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62172409), CCF-Baidu Open Fund, and Tencent Wechat Rhino-Bird Focused Research Program Research.

References

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

Ancuti, C. O., Ancuti, C., Sbert, M., and Timofte, R. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE international conference on image processing*, pp. 1014–1018, 2019.

Ancuti, C. O., Ancuti, C., and Timofte, R. 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*, June 2020.

Anwar, S. and Barnes, N. Real image denoising with feature attention. In *Proceedings of the IEEE International Conference on Computer Vision*, October 2019.

Cai, B., Xu, X., Jia, K., Qing, C., and Tao, D. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016.

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

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

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

Chen, S., Ye, T., Liu, Y., Liao, T., Ye, Y., and Chen, E. Msp-former: Multi-scale projection transformer for single image desnowing. *arXiv preprint arXiv:2207.05621*, 2022.

Chen, W.-T., Fang, H.-Y., Ding, J.-J., Tsai, C.-C., and Kuo, S.-Y. 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*, pp. 754–770, 2020.

Chen, W.-T., Fang, H.-Y., Hsieh, C.-L., Tsai, C.-C., Chen, I., Ding, J.-J., Kuo, S.-Y., 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*, pp. 4196–4205, 2021c.

Cheng, B., Li, J., Chen, Y., Zhang, S., and Zeng, T. Snow mask guided adaptive residual network for image snow removal. *arXiv preprint arXiv:2207.04754*, 2022.

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

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

Cui, Y., Tao, Y., Ren, W., and Knoll, A. Dual-domain attention for image deblurring. In *Association for the Advancement of Artificial Intelligence*, 2023b.

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

Dong, J. and Pan, J. Physics-based feature dehazing networks. In *Proceedings of the European Conference on Computer Vision*, pp. 188–204, 2020.

Engin, D., Genc, A., and Kemal Ekenel, H. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2018.

Fu, X., Huang, J., Ding, X., Liao, Y., and Paisley, J. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017a.

Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., and Paisley, J. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017b.

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

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

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

Hong, M., Xie, Y., Li, C., and Qu, Y. Distilling image dehazing with heterogeneous task imitation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

Jiang, K., Wang, Z., Yi, P., Chen, C., Huang, B., Luo, Y., Ma, J., and Jiang, J. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

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

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

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

Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., and Wang, Z. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1): 492–505, 2018a.

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

Li, R., Tan, R. T., and Cheong, L.-F. All in one bad weather removal using architectural search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

Li, X., Wu, J., Lin, Z., Liu, H., and Zha, H. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European Conference on Computer Vision*, September 2018b.

Li, Y., Tan, R. T., Guo, X., Lu, J., and Brown, M. S. Rain streak removal using layer priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

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

Liang, Y., Anwar, S., and Liu, Y. Drt: A lightweight single image deraining recursive transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 589–598, June 2022.

Lim, Y., Bliesener, Y., Narayanan, S., and Nayak, K. S. Deblurring for spiral real-time mri using convolutional neural networks. *Magnetic Resonance in Medicine*, 84 (6):3438–3452, 2020.

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

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

Liu, Y.-F., Jaw, D.-W., Huang, S.-C., and Hwang, J.-N. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6): 3064–3073, 2018a.

Liu, Y.-F., Jaw, D.-W., Huang, S.-C., and Hwang, J.-N. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6): 3064–3073, 2018b.

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

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

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

Park, N. and Kim, S. How do vision transformers work? In *International Conference on Learning Representations*, 2022.

Qin, X., Wang, Z., Bai, Y., Xie, X., and Jia, H. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11908–11915, 2020.

Qin, Z., Zhang, P., Wu, F., and Li, X. Fcanet: Frequency channel attention networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 783–792, October 2021.

Rasti, B., Chang, Y., Dalsasso, E., Denis, L., and Ghamisi, P. Image restoration for remote sensing: Overview and toolbox. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):201–230, 2021.

Ren, D., Zuo, W., Hu, Q., Zhu, P., and Meng, D. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

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

Rim, J., Kim, G., Kim, J., Lee, J., Lee, S., and Cho, S. Realistic blur synthesis for learning image deblurring. In *Proceedings of the European Conference on Computer Vision*, pp. 487–503, 2022.

Ruan, L., Chen, B., Li, J., and Lam, M. Learning to deblur using light field generated and real defocus images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 16304–16313, June 2022.

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

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

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

Tao, X., Gao, H., Shen, X., Wang, J., and Jia, J. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

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

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

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

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

Wei, W., Meng, D., Zhao, Q., Xu, Z., and Wu, Y. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

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

Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., and Ma, L. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10551–10560, June 2021.

Yang, W., Tan, R. T., Feng, J., Liu, J., Guo, Z., and Yan, S. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.

Yasarla, R. and Patel, V. M. 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*, June 2019.

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

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

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

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

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

Zang, S., Ding, M., Smith, D., Tyler, P., Rakotoarivelo, T., and Kaafar, M. A. 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*, 14(2):103–111, 2019.

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

Zhang, H., Sindagi, V., and Patel, V. M. Image de-raining using a conditional generative adversarial network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3943–3956, 2019b.

Zhang, K., Luo, W., Zhong, Y., Ma, L., Stenger, B., Liu, W., and Li, H. Deblurring by realistic blurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2020.

Zhang, K., Ren, W., Luo, W., Lai, W.-S., Stenger, B., Yang, M.-H., and Li, H. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., and Fu, Y. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, September 2018.

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

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

A. Datasets and Experimental Details

In this section, we introduce more details of the used datasets and training settings for five image restoration tasks. The summary of the datasets are shown in Table 14.

Single-Image Defocus Deblurring. Consistent with previous methods (Zamir et al., 2022; Ruan et al., 2022; Mao et al., 2021), we use a lately proposed dataset DPDD (Abuolaim & Brown, 2020) for evaluation. DPDD contains 500 indoor/outdoor scenes that are captured by a DSLR camera. In each scene, there are four images, termed as left, right and center views, and an all-in-focus ground-truth image. DPDD is split into training, validation, and testing subsets with 350, 74, and 76 scenes, respectively. The model take the center view images as input and predicts the sharp image. The training strategy follows that of (Ruan et al., 2022).

Image Dehazing. We use a synthetic dataset RESIDE (Li et al., 2018a) and two real-world datasets, Dense-Haze (Ancuti et al., 2019) and NH-HAZE (Ancuti et al., 2020), to evaluate IRNeXt. The RESIDE dataset consists of two training subsets: Indoor Training Set (ITS) that contains 13,990 hazy images produced from 1,399 sharp images, and Outdoor Training Set (OTS) that contains 313,950 hazy images generated from 8,970 clean images. In addition, RESIDE includes a Synthetic Objective Testing Set (SOTS), which consists of 500 indoor and 500 outdoor hazy images. We evaluate the ITS-trained and OTS-trained models on the corresponding SOTS datasets. For ITS and OTS, our models are trained for 300 and 30 epochs, respectively. And the batch size is set as 4 and 8, respectively. In addition, we use two real-world datasets to evaluate the robustness of IRNeXt in more challenging real-world scenarios. Both datasets contain 55 image pairs. NH-HAZE comprises nonhomogeneous hazy images while Dense-Haze contains dense and homogeneous hazy images. For these two datasets, our models are trained on 800×1200 patches with the batch size as 2 and the initial learning rate as $2e^{-4}$. Our models are trained for 5,000 epochs following the method (Guo et al., 2022).

Image Deraining. Following recent methods (Tu et al., 2022; Zamir et al., 2022; 2021), we utilize a composite training set to train our model, which consists of 13,712 clean and rainy image pairs collected from several datasets (Fu et al., 2017b; Yang et al., 2017; Zhang et al., 2019b; Li et al., 2016). The model is trained for 300 epochs.

Image Desnowing. For desnowing, we use three datasets for evaluation: SRRS (Chen et al., 2020), CSD (Chen et al., 2021c), and Snow100K (Liu et al., 2018a). The dataset settings are consistent with algorithms (Chen et al., 2022; 2020) where we randomly choose 2,000 images from each test set for evaluation. All models are trained for 800 epochs.

Image Motion Deblurring. Follow previous methods (Cho et al., 2021; Zamir et al., 2022; Tu et al., 2022; Wang et al., 2022), we train our model on the GoPro (Nah et al., 2017) dataset. It consists of 2,103 and 1,111 blurry/sharp image pairs for training and testing. The model is trained for 3000 epochs. We further evaluate the performance of our model on the real-world dataset RSBlur (Rim et al., 2022). It consist of 8,878 and 3,360 blurry/sharp image pairs for training and evaluation. Our model is trained for 710 epochs.

Method	Baseline	Groups				MSM	
		2	4	8	16	1	2
PSNR	31.23	31.55	31.57	31.64	31.58	31.64	31.75
Params/M	6.90	8.48	8.51	8.56	8.65	8.56	10.21

Table 11. Ablation studies of the number of groups in LAM and the number of MSM in CNNBlock on GoPro (Nah et al., 2017).

Method	Baseline	1 MSM	2 MSM
PSNR	31.33	38.58	40.23

Table 12. Ablation studies of the number of MSM in CNNBlock on the RESIDE-Indoor (Li et al., 2018a) dataset.

Blocks	12	14	16
PSNR	33.07	33.16	33.19
FLOPs/G	100.25	114.79	129.33
Params/M	11.66	13.21	14.76

Table 13. Ablation studies of the number of residual blocks in CNNBlock on the GoPro (Nah et al., 2017) dataset. The number includes the last one involving our MSM.

B. More Ablation Studies

We provide more ablation studies on the GoPro (Nah et al., 2017) and RESIDE-Indoor (Li et al., 2018a) datasets. The experimental configurations on GoPro are consistent with that of the ablation study in the main text. The experimental settings on RESIDE-Indoor follow the final model in Table 2, except that n is set to 0 in Figure 3 (c).

The number of groups in LAM. To study the impact of the diversity of attention weights in LAM, we conduct experiments by varying the number of groups. The results are presented in Table 11. Generally, as we increase the number of groups, the performance improves. However, it saturates at group 8, which is probably caused by overfitting.

The number of MSM. To further evaluate the effectiveness of MSM, we inject MSM into the last two residual blocks of each CNNBlock. The results on GoPro and RESIDE-indoor datasets are shown in Table 11 and Table 12, respectively. We can see that, deploying two MSM leads to the performance improvement on both datasets. To achieve a better

Task	Dataset	Training Set	Test Set
Defocus Deblurring	DPDD (Abuolaim & Brown, 2020)	350	76
Dehazing	RESIDE/ITS (Li et al., 2018a)	13990	500
	RESIDE/OTS (Li et al., 2018a)	313950	500
	Dense-Haze (Ancuti et al., 2019)	50	5
	NH-HAZE (Ancuti et al., 2020)	50	5
Draining	Rain14000 (Fu et al., 2017b)	11200	2800
	Rain1800 (Yang et al., 2017)	1800	0
	Rain800 (Zhang et al., 2019b)	700	100
	Rain100H (Yang et al., 2017)	0	100
	Rain100L (Yang et al., 2017)	0	100
	Rain12 (Li et al., 2016)	12	0
Desnowing	CSD(Chen et al., 2021c)	8000	2000
	SRRS (Chen et al., 2020)	15005	15005
	Snow100K (Liu et al., 2018a)	50000	50000
Motion Deblurring	GoPro (Nah et al., 2017)	2103	1111
	RSBlur (Rim et al., 2022)	8878	3360

Table 14. Details of the used datasets for five image restoration tasks.

balance between accuracy and computational overhead, we only insert one MSM into each CNNBlock.

The number of residual blocks in CNNBlock. We provide more model versions on the GoPro dataset for motion deblurring by varying n. The results are shown in Table 13. With 12 residual blocks (including the one involving our MSM), the model receives a performance gain of 0.15 dB PSNR over Restormer (Zamir et al., 2022). When we set $n = 16$, the accuracy increases to 33.19 dB. To compete with other algorithms in terms of performance and complexity, we finally choose $n = 14$ in our IRNeXt in Table 6.

C. More Details of the Low-Pass Filter

In Eq. 4, we define the low-pass filter as $\mathbf{A}_l = \frac{1}{K^2}\mathbf{E}$. In this section, we provide more details of this operation. We start from the 2D discrete Fourier transform:

$$F(u,v) = \frac{1}{K \cdot K} \sum_{x=0}^{K-1} \sum_{y=0}^{K-1} f(x,y) e^{-j2\pi(\frac{ux}{K} + \frac{vy}{K})} \quad (11)$$

where F and f denote the spectral and spatial features, respectively. j is the imaginary unit. $K \times K$ represents the region for Fourier transform. Since our low-pass filter is defined as a kind of filter that only retains the direct-current component, similar to (Qin et al., 2021), we suppose u and v in Eq. 11 are equal to 0, and then we have:

$$F(0,0) = \frac{1}{K \cdot K} \sum_{x=0}^{K-1} \sum_{y=0}^{K-1} f(x,y) e^{-j2\pi(\frac{0 \cdot x}{K} + \frac{0 \cdot y}{K})} \quad (12)$$

$$= \frac{1}{K \cdot K} \sum_{x=0}^{K-1} \sum_{y=0}^{K-1} f(x,y) \quad (13)$$

In our case, where the convolution operation is used for aggregation, for each pixel, we can obtain its low-frequency component based on Eq. 13:

$$\hat{\mathbf{X}}_{h,w}^{(0,0)} = \frac{1}{K \cdot K} \sum_{x=0}^{K-1} \sum_{y=0}^{K-1} 1 \cdot \mathbf{X}_{h-\lfloor \frac{K}{2} \rfloor + x, w-\lfloor \frac{K}{2} \rfloor + y} \quad (14)$$

where h and w denote spatial pixels of the feature. Our aggregation is performed on each region of size $K \times K$. $\hat{\mathbf{X}}_{h,w}^{(0,0)}$ means that for each region centered at pixel (h, w), we preserve the direct-current component. Comparing Eq.14 with Eq. 7, we can obtain the low-pass filter as $\mathbf{A}_l = \frac{1}{K^2}\mathbf{E}$.

D. Visualization

We visualize intermediate feature maps to demonstrate the effectiveness of our MSM and FM. We plot results of MSM by using different numbers of branches, as illustrated in the first three maps of Figure 9. The quantitative results are shown in Table 8. The two-branch version uses 4 and 2 as downsampling rates. We can see that using more branches generates sharper feature map. We further plot features for

Figure 9. Visualization of intermediate feature maps. The first three feature maps exhibit the difference of using different numbers of branches in MSM. The last three maps show the difference between branches with different downsampling rates.

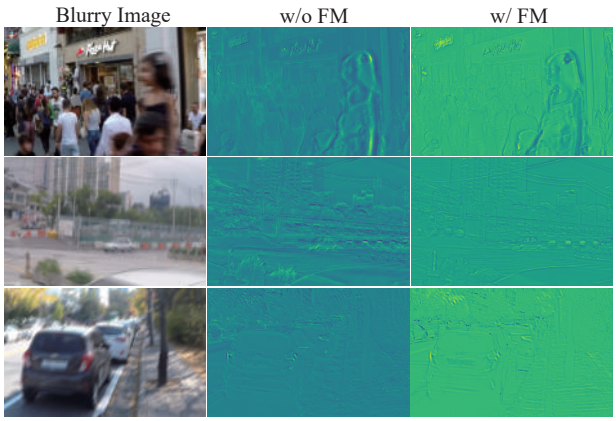

Figure 10. Visualization of intermediate feature maps of models with and without FM.

each branch in the three-branch version, which are obtained before addition in MSM. The last three maps show that the model restores sharp feature progressively, demonstrating the effectiveness of the coarse-to-fine mechanism.

Furthermore, we plot resulting feature maps of MSM for models with and without FM in Figure 10. The quantitative results are presented in Table 7c and Table 7d. As can be seen, equipped with FM, the model recovers more high-frequency signals than that without FM.

E. Additional Visual Results

In this part, we provide additional visual results for five image restoration tasks, organized as follows:

- Image deraining: Figure 11
- Image deblurring: Figure 12
- Image desnowing: Figure 13
- Image dehazing: Figure 14 and Figure 15

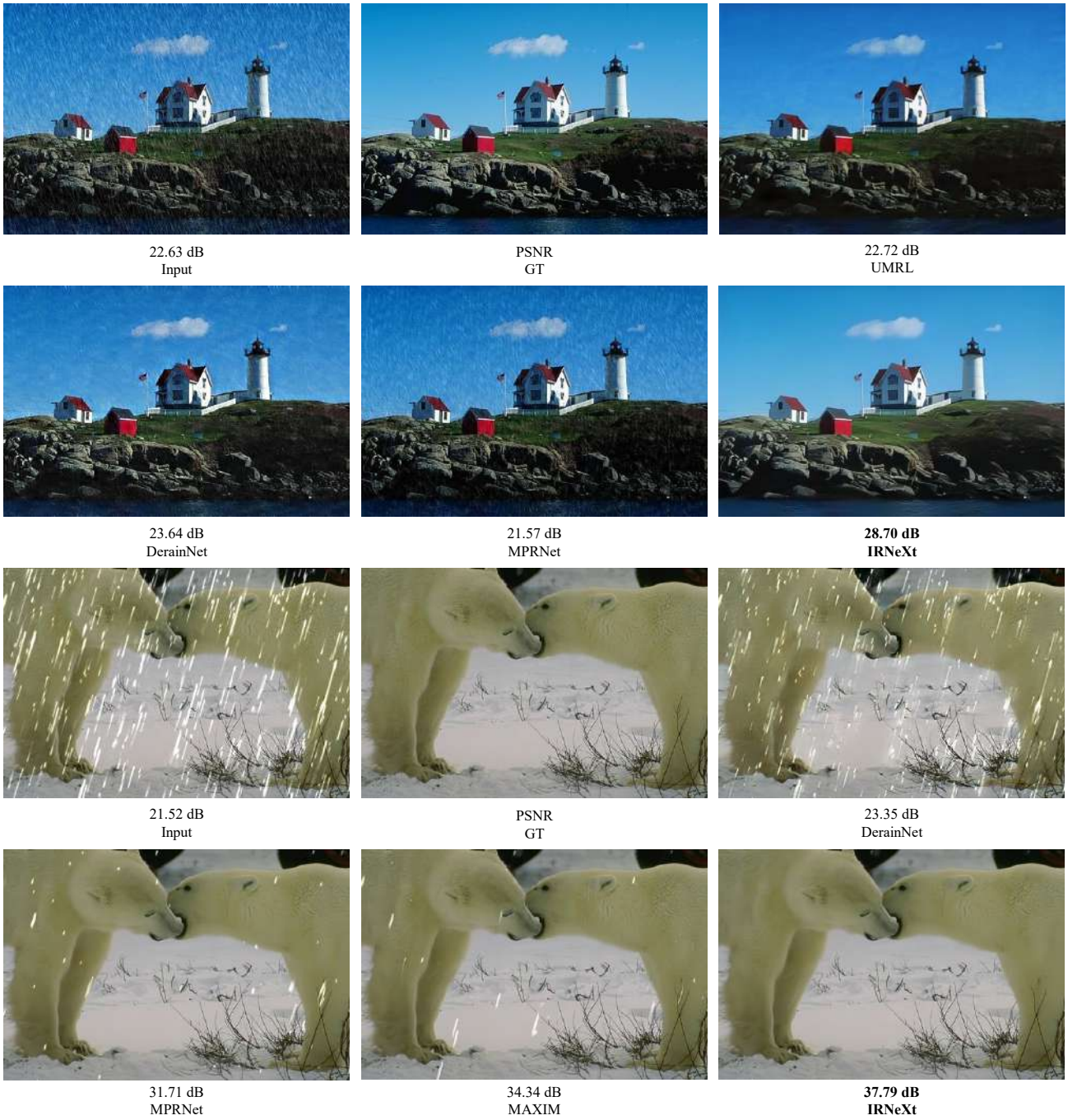

Figure 11. Image deraining comparisons. The top and bottom images are obtained from the Test100 (Zhang et al., 2019b) and Rain100L (Yang et al., 2017) datasets, respectively.

Figure 12. **Top**: Image motion deblurring comparisons on the real-world dataset RSBlur (Rim et al., 2022). **Bottom**: Single-image defocus deblurring comparisons on the DPDD (Abuolaim & Brown, 2020) dataset.

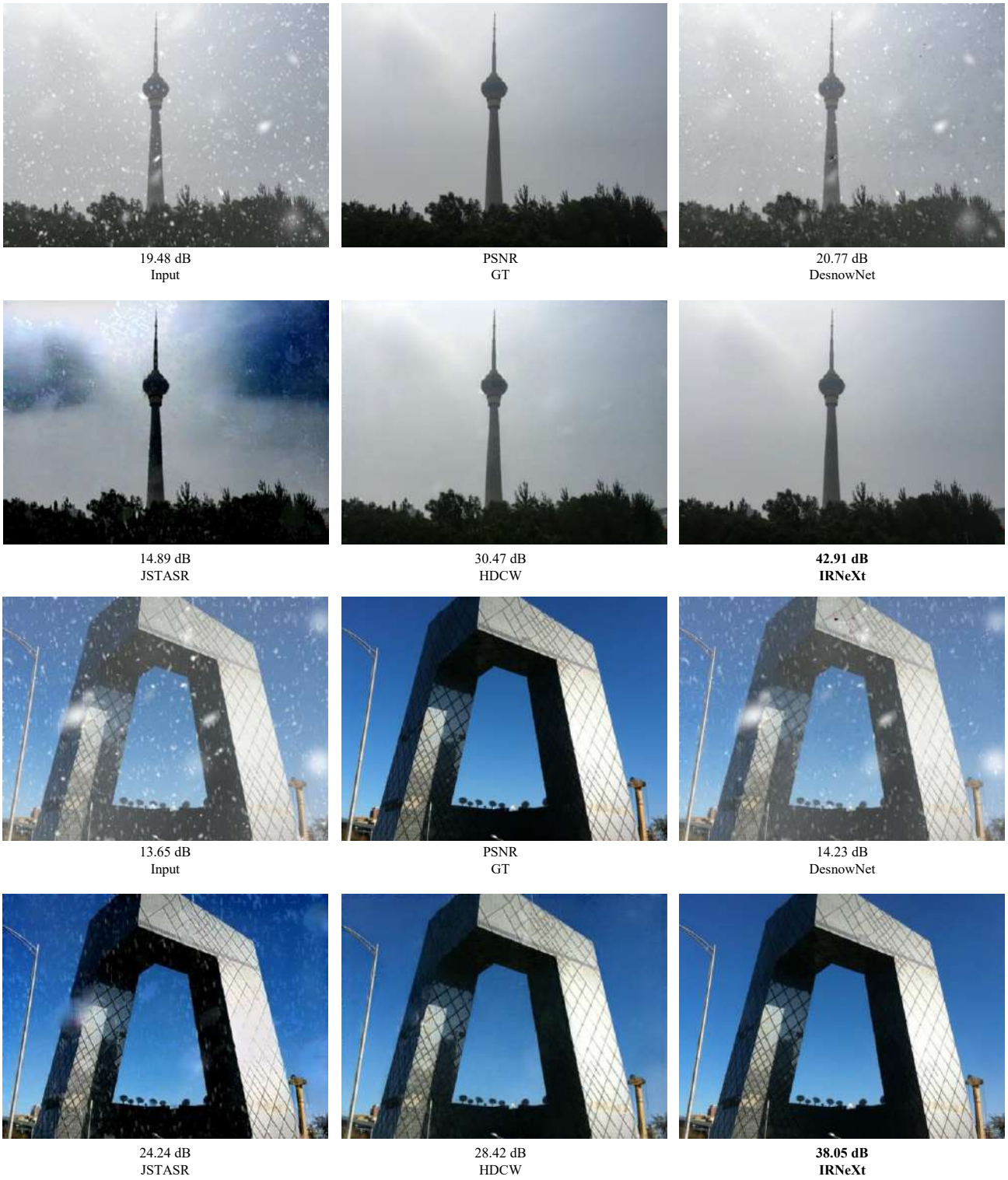

Figure 13. Image desnowing comparisons on the CSD (Chen et al., 2021c) dataset.

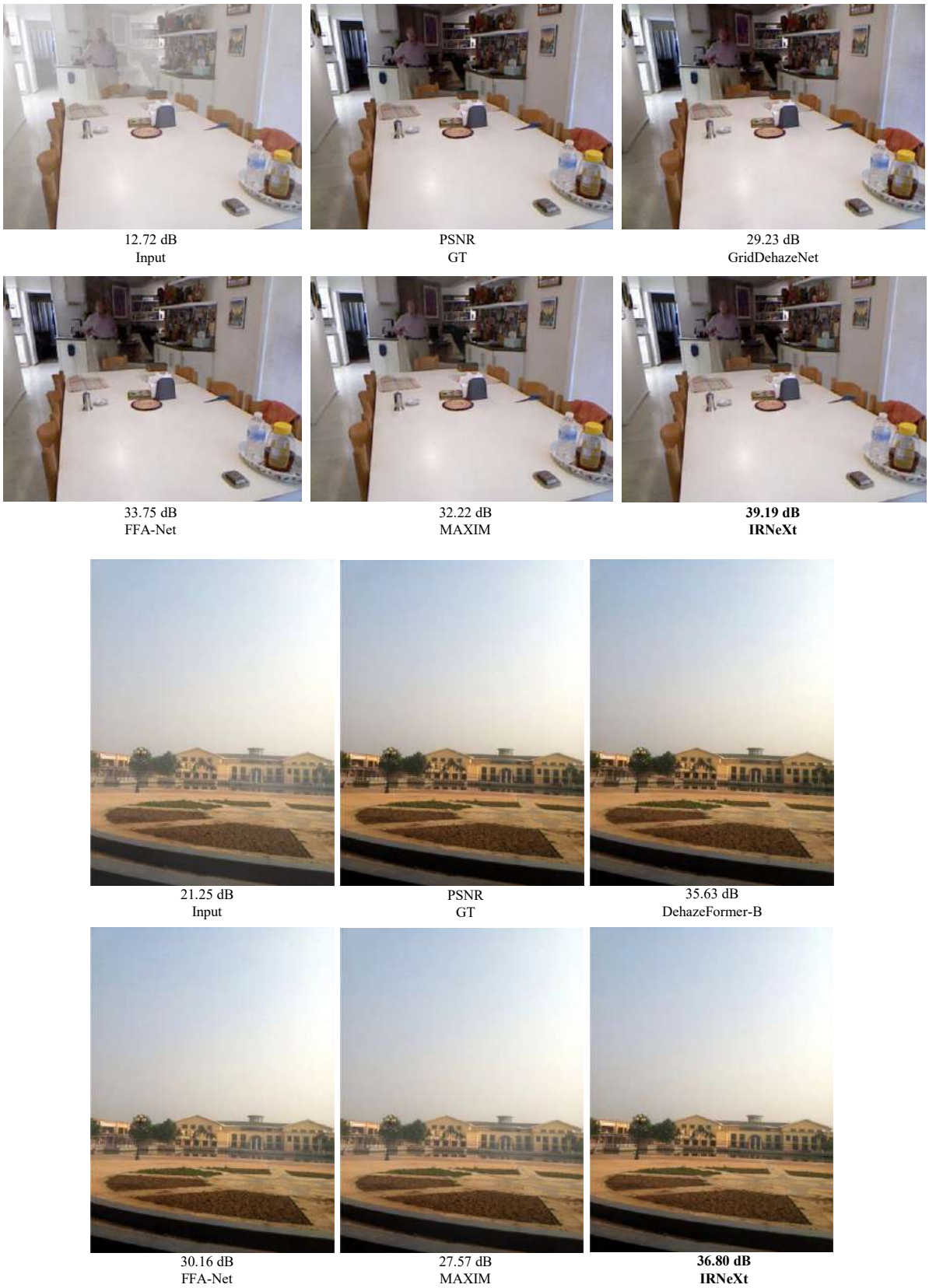

Figure 14. Image dehazing comparisons. The top and bottom images are obtained from the SOTS-Indoor (Li et al., 2018a) and SOTS-Outdoor (Li et al., 2018a) datasets, respectively.

Figure 15. Image dehazing comparisons. The top and bottom images are obtained from the Dense-Haze (Ancuti et al., 2019) and NH-HAZE (Ancuti et al., 2020) datasets, respectively.