Omni-Kernel Network for Image Restoration

Yuning Cui[1], **Wenqi Ren**[2*], **Alois Knoll**[1]

[1]Technical University of Munich
[2]Shenzhen Campus of Sun Yat-sen University
{yuning.cui, knoll}@in.tum.de, renwq3@mail.sysu.edu.cn

Abstract

Image restoration aims to reconstruct a high-quality image from a degraded low-quality observation. Recently, Transformer models have achieved promising performance on image restoration tasks due to their powerful ability to model long-range dependencies. However, the quadratically growing complexity with respect to the input size makes them inapplicable to practical applications. In this paper, we develop an efficient convolutional network for image restoration by enhancing multi-scale representation learning. To this end, we propose an omni-kernel module that consists of three branches, *i.e.,* global, large, and local branches, to learn global-to-local feature representations efficiently. Specifically, the global branch achieves a global perceptive field via the dual-domain channel attention and frequency-gated mechanism. Furthermore, to provide multi-grained receptive fields, the large branch is formulated via different shapes of depth-wise convolutions with unusually large kernel sizes. Moreover, we complement local information using a point-wise depth-wise convolution. Finally, the proposed network, dubbed OKNet, is established by inserting the omni-kernel module into the bottleneck position for efficiency. Extensive experiments demonstrate that our network achieves state-of-the-art performance on 11 benchmark datasets for three representative image restoration tasks, including image dehazing, image desnowing, and image defocus deblurring. The code is available at https://github.com/c-yn/OKNet.

Introduction

Image restoration aims to restore a sharp image from its low-quality counterpart, which suffers from degradations such as haze, snowflakes, and blur. To deal with this longstanding ill-posed problem, conventional approaches utilized various hand-crafted features and assumptions to restrict solution spaces. However, these methods are not applicable to more challenging real-world scenarios (Zhang et al. 2022).

In recent years, convolutional neural networks (CNNs) have achieved superior performance over traditional algorithms on image restoration tasks by learning generalizable priors from collected large-scale datasets (Chen et al. 2019). To further improve performance, many advanced functional units have been developed or borrowed from other domains for image restoration, such as the encoder-decoder architecture (Cho et al. 2021), residual connection (Mao et al. 2021), and attention mechanisms (Cui et al. 2023c). More recently, Transformer models have been introduced into image restoration and significantly advanced the state-of-the-art performance (Zamir et al. 2022). Nonetheless, the complexity of the core component, self-attention, grows quadratically with respect to the input size, making these methods unsuitable for practical applications.

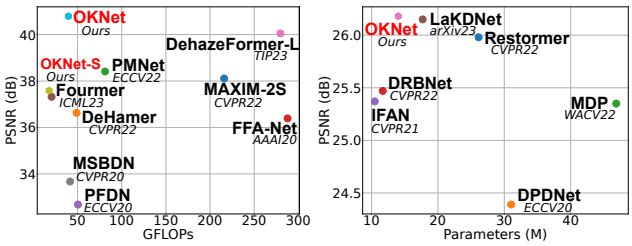

Figure 1: (a) FLOPs *vs.* PSNR on the SOTS-Indoor (Li et al. 2018) dataset for image dehazing. (b) The number of parameters *vs.* PSNR on the DPDD (Abuolaim and Brown 2020) dataset for image defocus deblurring. Our network achieves a better trade-off between performance and computation overhead over other state-of-the-art algorithms.

In contrast to the convolution operator that has limited receptive fields, Transformer models conduct global or large window-based self-attention, enabling networks to obtain large receptive fields. Inspired by this mechanism, a few recent works on CNNs strike back by designing efficient CNN frameworks with large kernels, such as 31×31 in RepLKNet (Ding et al. 2022) and 51×51 in SLaK (Liu et al. 2023). In the context of image restoration, LKDNet (Luo et al. 2022) decomposes a 21×21 convolution into a smaller depth-wise convolution and a depth-wise dilated convolution for image dehazing. LaKDNet (Ruan et al. 2023) leverages large kernel size convolutions (*e.g.,* 9×9) followed by point-wise convolutions to obtain large effective receptive fields for image deblurring. MAN (Wang et al. 2022b) decomposes a large kernel size convolution into three components, *i.e.,* a depth-wise convolution, a depth-wise dilated

*Corresponding author

convolution, and a point-wise convolution. However, the receptive fields produced by these methods are still limited, and they do not provide global receptive fields.

In this paper, we explore the potential of unusually large kernel size convolutions for image restoration by using a 63×63 depth-wise convolution. Furthermore, we utilize large strip-based convolutions to further enhance representation learning for high-quality image reconstruction. To restrain computation overhead brought by these large convolutions, we deploy them only in the bottleneck. Moreover, we resort to the dual-domain channel attention and frequency-gated mechanism to provide global receptive fields. In addition to pursuing large receptive fields, we also utilize a 1×1 depth-wise convolution to complement local information for small-size degradations. Finally, the proposed omni-kernel module (OKM) is formed by organizing the above designs in parallel such that the network possesses the ability to handle multi-scale degradations.

Equipped with OKM in the bottleneck, our simple convolutional network achieves state-of-the-art performance on 11 different datasets for three representative image restoration tasks. More concretely, OKNet significantly outperforms the recent Transformer model Fourmer (Zhou et al. 2023) by 3.47 dB PSNR with comparable complexity on the SOTS-Indoor (Li et al. 2018) dataset, as illustrated in Figure 1 (a). For single-image defocus deblurring, our model achieves a performance gain of 0.2 dB PSNR over the strong Transformer model Restormer (Zamir et al. 2022) in the combined category of the DPDD (Abuolaim and Brown 2020) dataset with 46% fewer parameters. Furthermore, the proposed model also represents the strong capability on the image desnowing task, outperforming the recent algorithm IRNeXt (Cui et al. 2023c) by 0.7 dB PSNR on the CSD (Chen et al. 2021b) dataset. Overall, the contributions of this paper can be summarized as follows:

- We present an omni-kernel module that is capable of efficiently capturing multi-scale receptive fields for image restoration, among which large-scale information is modulated via the dual-domain processing and different shapes of large kernel size depth-wise convolutions.
- Extensive experiments on 11 widely used benchmark datasets demonstrate that the proposed model, namely OKNet, achieves state-of-the-art performance on three representative image restoration tasks, *i.e.,* image defocus deblurring, image dehazing, and image desnowing.

Related Works

Image Restoration

As a longstanding problem, image restoration aims to reconstruct a clean image from its degraded counterpart, playing an important role in many scenarios, such as surveillance, self-driving technology, remote sensing, and medical imaging. Due to its highly ill-posed property, many conventional algorithms have been developed mainly based on assumptions and hand-crafted features, which are not applicable to more challenging practical applications.

In recent years, deep learning methods have achieved notably superior performance over traditional competitors by learning generalizable priors from large-scale datasets. These approaches can be roughly divided into two categories: CNN-based and Transformer-based methods. CNN-based methods have dominated image restoration for many years by designing or borrowing advanced functional units from other domains (Cui et al. 2023b; Cui and Knoll 2023). For instance, FFA-Net (Qin et al. 2020) utilizes the channel attention and pixel attention modules to treat channel-wise and pixel-wise features unequally for uneven haze distribution. SDWNet (Zou et al. 2021) leverages multiple dilated convolutions with different dilated rates in parallel to obtain large receptive fields. SFNet (Cui et al. 2023d) uses a dynamic selective frequency module to select the most informative frequency to recover. To model long-range dependencies more efficiently, Transformer (Vaswani et al. 2017) has been introduced into image restoration (Chen et al. 2021a; Liang et al. 2021). To improve the efficiency of self-attention, the common strategies are restricting self-attention regions (Wang et al. 2022c; Tsai et al. 2022) and switching self-attention from the spatial dimension to the channel dimension (Zamir et al. 2022). Despite these efforts, Transformer models are still expensive for practical applications and sacrifice the long-range signals modeling capability. Furthermore, Transformer models cannot capture multi-scale receptive fields. In this paper, we present an efficient convolutional network that is capable of learning multi-scale representations.

Large Kernel Network

Recently, inspired by the plausible reason behind the success of Transformer, *i.e.,* the long-range dependencies modeling ability, CNN-based methods strike back by using large kernel convolutions. For instance, RepLKNet (Ding et al. 2022) achieves a 31×31 kernel following several guidelines for designing large convolutions, greatly closing the performance gap between CNNs and Transformer models. SLaK (Liu et al. 2023) leverages sparse factorized 51×51 kernels to confront Transformer methods. In the realm of image restoration, LaKDNet (Ruan et al. 2023) enlarges the effective receptive field via combinations of large kernel (9×9) depth-wise convolutions and point-wise convolutions. MAN (Wang et al. 2022b) presents the large kernel attention by decomposing a large kernel convolution into three different kinds of convolutions. LKD-Net (Luo et al. 2022) decomposes a depth-wise convolution into a smaller depth-wise convolution and a depth-wise dilated convolution. Our method is different from above image restoration algorithms in fourfold: **(a)** we explore the potential of unusually large kernel size convolutions for image restoration, *i.e.,* 63×63; **(b)** apart from the regular square depth-wise convolution, we use strip-based versions in different directions to provide different shapes of receptive fields for high-quality image reconstruction; **(c)** we offer full-size receptive fields via the dual-domain processing; **(d)** we complement local information via an extremely simple 1×1 depth-wise convolution. Deploying the proposed method only in the bottleneck, our efficient OKNet can perform on par with or better than state-of-the-art Transformer models.

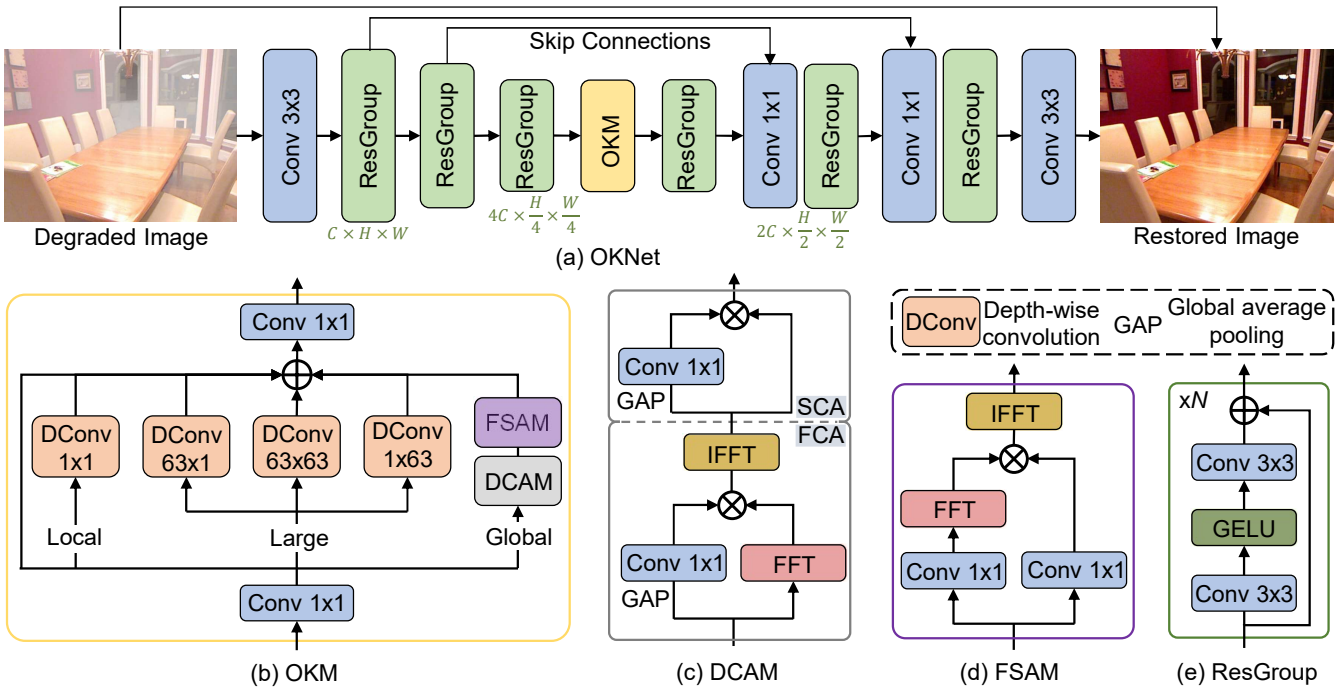

Figure 2: The architecture of OKNet. FFT and IFFT denote fast Fourier transform and its inverse operation, respectively.

Methodology

In this section, we first describe the overall pipeline of the proposed OKNet. Then, we delineate the architectural details of our omni-kernel module (OKM). Finally, we present the used loss functions for the training stage.

Overall Pipeline

The overall pipeline is illustrated in Figure 2 (a). As shown, OKNet adopts an encoder-decoder architecture, which consists of three scales in both the encoder and decoder stages. ResGroup is composed of multiple residual blocks, each including two 3×3 convolutions with the GELU (Hendrycks and Gimpel 2016) nonlinearity in between. OKM is only inserted into the bottleneck position, where features have the lowest resolution, for saving computation overhead.

Given an input degraded image $I \in \mathbb{R}^{3\times H\times W}$, we first leverage a 3×3 convolution to project the image into embedding features of size $C \times H \times W$, where C denotes the number of channels, and $H \times W$ specifies the spatial location of pixels. Next, the resulting features are fed into the encoder stage to extract in-depth representations. The downsampling operation is implemented by a strided convolution (*kernel*=3, *stride*=2), which expands the number of channels while reducing the spatial dimension. After being processed by the proposed OKM, the features pass through the decoder network to restore high-resolution representations. During this process, the decoder features are concatenated with the encoder features to assist in recovery, followed by a 1×1 convolution to reduce channels by half. The upsampling layer is accomplished by a transposed convolution (*kernel*=4, *stride*=2) to enlarge the spatial dimension and halve the number of channels. Finally, a 3×3 convolution is used to yield the learned residual image, to which the input image is added to produce the final restored output. Next, we introduce the proposed OKM in detail.

Omni-Kernel Module (OKM)

The schematic diagram of OKM is illustrated in Figure 2 (b). Given input features $X \in \mathbb{R}^{C\times H\times W}$, after being processed by a 1×1 convolution, the features are fed into three branches, *i.e.,* local branch, large branch and global branch, to enhance multi-scale representations. The results of the three branches are then fused by addition and modulated by another 1×1 convolution. In the following, we introduce the inside components of each branch.

Large Branch In this branch, we apply a cheap depth-wise convolution of kernel size $K \times K$ to pursue large receptive fields. In addition to the regular depth-wise convolution, inspired by strip-based self-attention (Dong et al. 2022; Tsai et al. 2022; Li et al. 2023), we also employ $1 \times K$ and $K \times 1$ depth-wise convolutions in parallel to the square one to harvest strip-shaped contextual information. To avoid introducing a large amount of computation overhead caused by large kernel size convolutions, we place the module in the bottleneck position. We then explore the possibility of using extremely large convolutions for image restoration by progressively increasing K. The experimental results are shown in Figure 3. Generally speaking, the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) metrics increase as we enlarge the kernel size from $K = 3$ to $K = 63$. The placed location of our module allows us to adopt an unusually large kernel size for capturing large-scale different

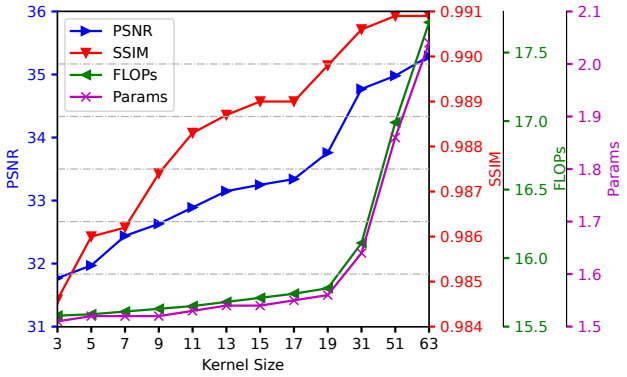

Figure 3: Experimental results on different kernel sizes of convolutions in the large branch.

shapes of receptive fields with few introduced parameters and low complexity. We finally choose $K = 63$ in the large branch for better performance.

Global Branch OKNet is mainly trained on cropped $3 \times 256 \times 256$ image patches and the bottleneck features have a spatial size of 64×64, and thus we adopt the largest odd kernel size in the large branch. However, during the inference stage, input degraded images are much larger than training patches. As a result, a 63×63 kernel is not capable of covering the global field. To alleviate this issue, we superadd the global modeling capability in the global branch by resorting to dual-domain processing. Specifically, the global branch consists of a dual-domain channel attention module (DCAM) and a frequency-based spatial attention module (FSAM). Next, we present these two modules successively.

Given input features $X_{Global} \in \mathbb{R}^{C \times H \times W}$, DCAM firstly applies frequency channel attention (FCA) to X_{Global} as:

$$X_{\text{FCA}} = \mathcal{IF}(\mathcal{F}(X_{Global}) \otimes W_{1\times1}^{\text{FCA}}(\text{GAP}(X_{Global}))) \tag{1}$$

where $\mathcal{F}$ and $\mathcal{IF}$ are fast Fourier transformer and its inverse operation, respectively; X_{FCA}, $W_{1\times1}$ and GAP denote the output of FCA, a 1×1 convolutional layer and global average pooling, respectively; $\otimes$ represents the element-wise multiplication operation. With the Fourier processing, the global features are refined effectively according to the spectral convolution theorem. After being modulated globally in the spectral domain, the resulting features are further fed into the spatial channel attention module (SCA), which can be formally expressed as:

$$X_{\text{DCAM}} = X_{\text{FCA}} \otimes W_{1\times1}^{\text{SCA}}(\text{GAP}(X_{\text{FCA}})) \tag{2}$$

where X_{DCAM} is the output of DCAM. DCAM only enhances dual-domain features at the channel-wise coarse granularity. Then, we apply the frequency-based attention module among the spatial dimension to refine the spectrum at a fine-grained level, which is formally expressed as:

$$X_{\text{FSAM}} = \mathcal{IF}(\mathcal{F}(W_{1\times1}^{1}(X_{\text{DCAM}})) \otimes W_{1\times1}^{2}(X_{\text{DCAM}})) \tag{3}$$

where X_{FSAM} is the result of FSAM. By doing this, the model attends to informative frequency components for high-quality image reconstruction.

Local Branch Inspired by the fact that local information plays an essential role in image restoration (Zamir et al. 2022; Wang et al. 2022c), in addition to the large and global branches that capture large-scale receptive fields, we also design an extremely simple yet effective local branch for local signals modulation by using a 1×1 depth-wise convolutional layer, as illustrated in Figure 2 (b). We demonstrate its effectiveness in Table 5.

Loss Function

To restore faithful high-quality images, a straightforward way is to make the content of the predicted image closer to that of the ground truth image:

$$\mathcal{L}_c = \|\hat{I} - Y\|_1 \tag{4}$$

where $\hat{I}$ and Y represent the predicted image and ground truth, respectively. In addition to spatial domain alignment, the proposed network also promotes frequency signal learning. Accordingly, we additionally apply the frequency domain L_1 loss (Cho et al. 2021) for training:

$$\mathcal{L}_f = \|\mathcal{F}(\hat{I}) - \mathcal{F}(Y)\|_1 \tag{5}$$

Finally, the overall loss function is given by:

$$\mathcal{L}_o = \frac{1}{E}(\mathcal{L}_c + \lambda \mathcal{L}_f) \tag{6}$$

where E represents the number of elements in the output, and λ is set to 0.1 for balancing dual-domain training.

Experiments

In this section, we conduct experiments on 11 different benchmark datasets to demonstrate the effectiveness of our network for three representative image restoration tasks: image dehazing, image defocus deblurring, and image desnowing. In the tables, the best and second best results are marked in **bold** and underlined.

Implementation Details

We train separate models for different datasets. According to the task complexity, we scale OKNet by setting different numbers of residual blocks in each ResGroup (Figure 2 (e)), *i.e.,* $N = 4$ for dehazing and desnowing, and $N = 16$ for deblurring. Additionally, we provide a small version for image dehazing, dubbed OKNet-S, by setting $N = 1$ to better compare with the recent algorithm (Zhou et al. 2023). Unless stated otherwise, the following hyperparameters are adopted. The models are trained using the Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The batch size is set to 8. The learning rate is initially set to $2e^{-4}$ and decreased to $1e^{-6}$ gradually using the cosine annealing decay strategy (Loshchilov and Hutter 2016). For data augmentation, the cropped patches of size 256×256 are randomly horizontally flipped with a probability of 0.5. FLOPs are measured on 256×256 patch size.

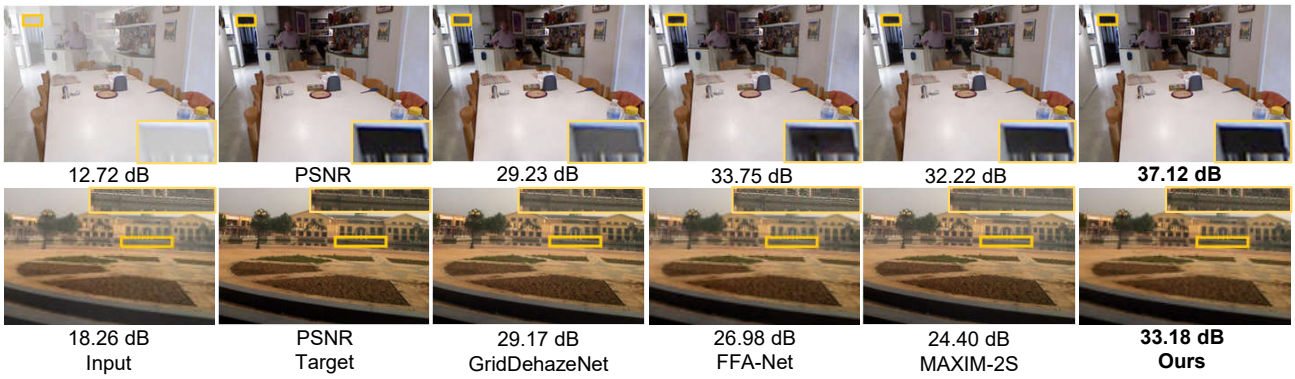

Figure 4: Image dehazing comparisons on the SOTS (Li et al. 2018) dataset. The top and bottom images are obtained from SOTS-Indoor and SOTS-Outdoor, respectively. Our results are produced by OKNet.

Method	SOTS-Indoor		SOTS-Outdoor		Dense-Haze		NH-HAZE		O-HAZE		I-Haze	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
GridDehazeNet	32.16	0.984	30.86	0.982	13.31	0.37	13.80	0.54	23.51	0.83	18.73	0.77
MSBDN	33.67	0.985	33.48	0.982	15.37	0.49	19.23	0.71	24.36	0.75	19.62	0.62
FFA-Net	36.39	0.989	33.57	0.984	14.39	0.45	19.87	0.69	22.12	0.77	19.72	0.73
PMNet	38.41	0.990	34.74	0.985	16.79	0.51	20.42	0.73	24.64	0.83	-	
MAXIM-2S	38.11	0.991	34.19	0.985	-		-		-		-	
DeHamer	36.63	0.988	35.18	0.986	16.62	0.56	**20.66**	0.68	-		-	
SDCE	-		-		16.85	0.60	20.42	0.74	24.92	0.84	20.81	0.82
DehazeFormer-L	40.05	**0.996**	-		-		-		-		-	
Fourmer	37.32	0.990	-		15.95	0.49	19.91	0.72	-		-	
OKNet-S	37.59	0.994	35.45	0.992	16.85	0.62	20.29	**0.80**	25.18	0.93	20.69	0.85
OKNet	**40.79**	**0.996**	**37.68**	**0.995**	**16.92**	**0.64**	20.48	**0.80**	**25.64**	**0.94**	**21.72**	**0.87**

Table 1: Image dehazing comparisons on the daytime synthetic and real-world datasets.

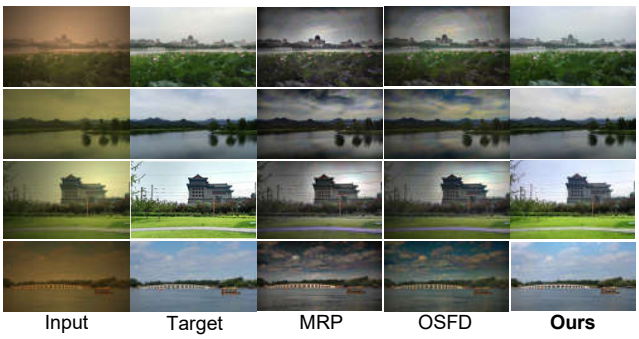

Figure 5: Nighttime image dehazing comparisons on the NHR (Zhang et al. 2020) dataset.

Method	GS	MRPF	MRP	OSFD	HCD	FocalNet	**OKNet**
PSNR	17.32	16.95	19.93	21.32	23.43	25.35	**27.92**
SSIM	0.629	0.667	0.777	0.804	0.953	0.969	**0.979**

Table 2: Nighttime image dehazing comparisons on the NHR (Zhang et al. 2020) dataset.

Experimental Results

Image Dehazing Results We conduct dehazing experiments on three kinds of datasets: daytime synthetic dataset (RESIDE (Li et al. 2018)), daytime real-world datasets (Dense-Haze (Ancuti et al. 2019), NH-HAZE (Ancuti, Ancuti, and Timofte 2020), O-Haze (Ancuti et al. 2018b), and I-Haze (Ancuti et al. 2018a)), and nighttime dataset (NHR (Zhang et al. 2020)). For daytime datasets, we compare our models with 9 representative state-of-the-art methods: GridDehazeNet (Liu et al. 2019), MSBDN (Dong et al. 2020), FFA-Net (Qin et al. 2020), PMNet (Ye et al. 2022), MAXIM-2S (Tu et al. 2022), DeHamer (Guo et al. 2022), SDCE (Zhu et al. 2023), DehazeFormer-L (Song et al. 2023), and Fourmer (Zhou et al. 2023). The results are presented in Table 1. Our OKNet achieves the best results on most metrics. Specifically, OKNet outperforms the expensive Transformer-based model DehazeFormer-L by 0.74 dB PSNR on the SOTS-Indoor (Li et al. 2018) dataset with only 14% FLOPs as shown in Figure 1 (a). Furthermore, our OKNet produces better performance on all real-world datasets over SDCE, which is elaborately designed for real-world scenarios. Moreover, our OKNet-S achieves a significant gain of 0.27 dB PSNR on SOTS-Indoor over the recent algorithm Fourmer with 13% fewer FLOPs. The visual comparisons on SOTS-Indoor and SOTS-Outdoor are illustrated in Figure 4. Our OKNet generates more faithful results than

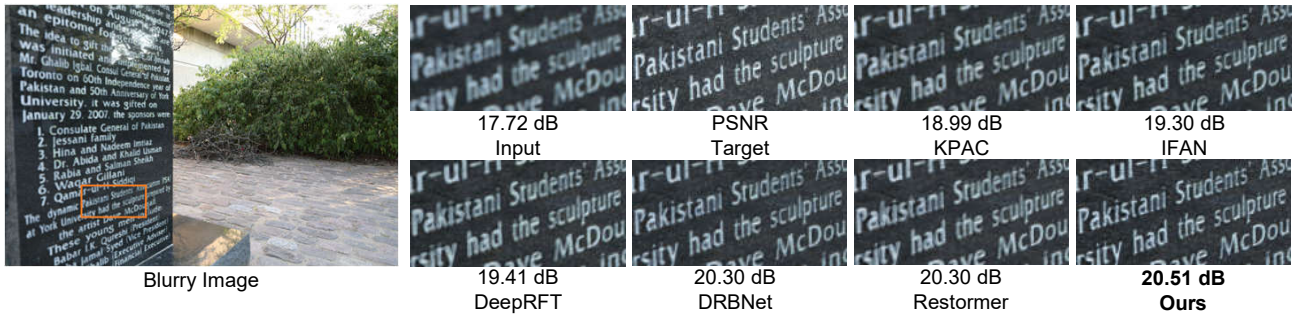

Figure 6: Image defocus deblurring comparisons on the DPDD (Abuolaim and Brown 2020) dataset.

Method	Indoor Scenes				Outdoor Scenes				Combined			
	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓
DPDNet	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	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227
DRBNet			-				-		25.73	0.791	-	0.183
IFAN	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217
MDP	28.02	0.841	0.027	-	22.82	0.690	0.052	-	25.35	0.763	0.040	0.303
Restormer	28.87	**0.882**	0.025	**0.145**	23.24	0.743	0.050	**0.209**	25.98	0.811	0.038	0.178
LaKDNet			-				-		26.15	0.810	-	**0.155**
OKNet	**28.99**	0.877	**0.024**	0.169	**23.51**	**0.751**	**0.049**	0.241	**26.18**	**0.812**	**0.037**	0.206

Table 3: Single-image defocus deblurring comparisons on the DPDD (Abuolaim and Brown 2020) dataset.

other competitive algorithms.

We further present comparisons on the nighttime image dehazing dataset NHR (Zhang et al. 2020) with 6 state-of-the-art methods: GS (Li, Tan, and Brown 2015), MRPF (Zhang et al. 2017), MRP (Zhang et al. 2017), OSFD (Zhang et al. 2020), HCD (Wang et al. 2022a), and FocalNet (Cui et al. 2023a). Table 2 shows that our method outperforms the recent FocalNet by 2.57 dB PSNR and 0.01 SSIM. Figure 5 illustrates that the results yielded by our network are closer to the ground-truth targets.

Image Defocus Deblurring Results We verify the effectiveness of the proposed network for single-image defocus deblurring using the widely used DPDD (Abuolaim and Brown 2020) dataset, and compare the results with 7 representative algorithms: DPDNet (Abuolaim and Brown 2020), KPAC (Son et al. 2021), DRBNet (Ruan et al. 2022), IFAN (Lee et al. 2021), MDP (Abuolaim, Afifi, and Brown 2022), Restormer (Zamir et al. 2022), and LaKDNet (Ruan et al. 2023). The results are shown in Table 3. As seen, our model achieves better performance over other methods on most metrics. Concretely, OKNet obtains a remarkable gain of 0.27 dB PSNR over the strong Transformer-based model Restormer in the outdoor scenes. Furthermore, compared with LaKDNet, which also uses large kernel convolutions, our model yields performance gains of 0.03 dB PSNR and 0.002 SSIM on the combined category with 21% fewer parameters, as illustrated in Figure 1 (b). The visual comparisons are shown in Figure 6. Our method generates a sharper and more visually-faithful result than other competitors.

Method	CSD		SRRS		Snow100K	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DesnowNet	20.13	0.81	20.38	0.84	30.50	0.94
All in One	26.31	0.87	24.98	0.88	26.07	0.88
JSTASR	27.96	0.88	25.82	0.89	23.12	0.86
HDCW-Net	29.06	0.91	27.78	0.92	31.54	**0.95**
TransWeather	31.76	0.93	28.29	0.92	31.82	0.93
FocalNet	37.18	**0.99**	31.34	**0.98**	33.53	**0.95**
IRNeXt	37.29	**0.99**	**31.91**	**0.98**	33.61	**0.95**
OKNet	**37.99**	**0.99**	31.70	**0.98**	**33.75**	**0.95**

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

Image Desnowing Results We evaluate the proposed model on three widely used datasets for image desnowing, including Snow100K (Liu et al. 2018), SRRS (Chen et al. 2020), and CSD (Chen et al. 2021b). We then compare our results with 8 state-of-the-art algorithms: DesnowNet (Liu et al. 2018), CycleGAN (Engin, Genc, and Kemal Ekenel 2018), All in One (Li, Tan, and Cheong 2020), JSTASR (Chen et al. 2020), HDCW-Net (Chen et al. 2021b), TransWeather (Valanarasu, Yasarla, and Patel 2022), FocalNet (Cui et al. 2023a), and IRNeXt (Cui et al. 2023c). Table 4 shows that the proposed network has a strong capability for snow removal. Specifically, OKNet outperforms the recent algorithm IRNeXt by 0.14 dB PSNR on the Snow100K dataset. On the more complicated CSD dataset, the advantage becomes greater, suggesting the effectiveness

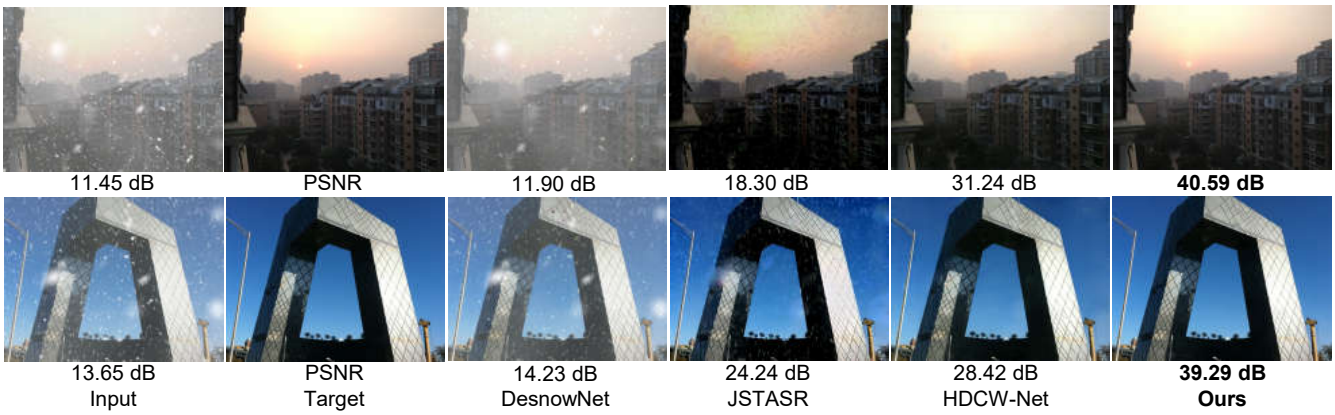

Figure 7: Image desnowing comparisons on the CSD (Chen et al. 2021b) dataset.

#	Baseline	Large Branch		Small Branch	Global Branch			PSNR	SSIM	Params/M	FLOPs/G
		Square Conv	Strip Conv		DCAM/FCA	DCAM/SCA	FSAM				
1	✓							31.32	0.98357	1.48	15.44
2	✓	✓						35.07	0.99082	2.02	17.65
3	✓	✓	✓					35.29	0.99088	2.04	17.72
4	✓	✓	✓	✓				35.48	0.99120	2.04	17.72
5	✓				✓			32.84	0.98748	1.49	15.44
6	✓	✓	✓	✓	✓			35.82	0.99151	2.05	17.72
7	✓					✓		32.35	0.98676	1.49	15.44
8	✓	✓	✓	✓	✓	✓		36.12	0.99188	2.07	17.72
9	✓						✓	33.32	0.98879	1.81	15.57
10	✓	✓	✓	✓	✓	✓	✓	36.48	0.99204	2.40	17.86

Table 5: Ablation studies for the proposed components.

of our method. Figure 7 shows that our results are more visually pleasing than others.

Ablation Studies

We perform ablation studies by training OKNet-S on the RESIDE-Indoor (Li et al. 2018) dataset for 300 epochs and evaluating on SOTS-Indoor (Li et al. 2018). The baseline model is obtained by removing OKM from our model.

We progressively add the designed large branch, small branch, and global branch to the baseline model. The results are shown in Table 5. The baseline model achieves 31.32 dB PSNR. The unusually large regular convolution leads to a significant gain of 3.75 dB over baseline, while the strip-based convolutions further improve the performance to 35.29 dB, demonstrating the effectiveness of capturing different shapes of receptive fields. The extremely simple small branch boosts the accuracy to 35.48 dB by enhancing local information. Finally, we investigate the efficacy of individual components in the global branch. FCA, SCA, and FSAM achieve performance gains of 1.52 dB, 1.03 dB, and 2 dB over the baseline model. The combination (#8) of FCA and SCA yields a higher score than only using FCA (#6), suggesting the compatibility of our designs. Equipped with FSAM, the full model produces the best result, which is 5.16 dB higher than that of the baseline model.

Conclusion

In this paper, we propose an efficient convolutional network, dubbed OKNet, which is capable of capturing multi-scale receptive fields. The core component, OKM, consists of three branches for modeling local, large, and global dependencies. The large branch is designed by exploring the unusually large regular and strip-based depth-wise convolutions for image restoration. The novel global branch utilizes dual-domain channel attention and frequency-based spatial attention for modulating global representations. Furthermore, the extremely lightweight local branch brings locality to the model. Inserting the simple yet effective OKM into the bottleneck, OKNet achieves state-of-the-art performance on 11 different datasets for three image restoration tasks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62322216, 62172409), Shenzhen Science and Technology Program (Grant No. JCYJ20220818102012025, RCYX20221008092849068), 2023 CCF-Tencent Rhino-Bird Young Faculty Open Research Fund and CCF-Zhejiang Lab Joint Innovation Fund.

References

Abuolaim, A.; Afifi, M.; and Brown, M. S. 2022. 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*, 1231–1239.

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

Ancuti, C.; Ancuti, C. O.; Timofte, R.; and De Vleeschouwer, C. 2018a. 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*, 620–631.

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

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

Ancuti, C. O.; Ancuti, C.; Timofte, R.; and De Vleeschouwer, C. 2018b. 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*, 754–762.

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

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021a. Pre-Trained Image Processing Transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

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

Chen, W.-T.; Fang, H.-Y.; Hsieh, C.-L.; Tsai, C.-C.; Chen, I.; Ding, J.-J.; Kuo, S.-Y.; et al. 2021b. 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*, 4196–4205.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking Coarse-To-Fine Approach in Single Image Deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, 4641–4650.

Cui, Y.; and Knoll, A. 2023. Exploring the potential of channel interactions for image restoration. *Knowledge-Based Systems*, 282: 111156.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023a. Focal Network for Image Restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, 13001–13011.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023b. Image Restoration Via Frequency Selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cui, Y.; Ren, W.; Yang, S.; Cao, X.; and Knoll, A. 2023c. IRNeXt: Rethinking Convolutional Network Design for Image Restoration. In *International Conference on Machine Learning*.

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

Ding, X.; Zhang, X.; Han, J.; and Ding, G. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11963–11975.

Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; and Yang, M.-H. 2020. Multi-Scale Boosted Dehazing Network With Dense Feature Fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; and Guo, B. 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12124–12134.

Engin, D.; Genc, A.; and Kemal Ekenel, H. 2018. Cycle-Dehaze: Enhanced CycleGAN for Single Image Dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Guo, C.-L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and Li, C. 2022. Image Dehazing Transformer with Transmission-Aware 3D Position Embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5812–5820.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

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

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

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2018. Benchmarking Single Image Dehazing and Beyond. *IEEE Transactions on Image Processing*.

Li, R.; Tan, R. T.; and Cheong, L.-F. 2020. All in One Bad Weather Removal Using Architectural Search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 18278–18289.

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

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

Liu, S.; Chen, T.; Chen, X.; Chen, X.; Xiao, Q.; Wu, B.; Kärkkainen, T.; Pechenizkiy, M.; Mocanu, D. C.; and Wang, Z. 2023. More ConvNets in the 2020s: Scaling up Kernels Beyond 51x51 using Sparsity. In *International Conference on Learning Representations*.

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

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

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

Luo, P.; Xiao, G.; Gao, X.; and Wu, S. 2022. LKD-net: Large kernel convolution network for single image dehazing. *arXiv preprint arXiv:2209.01788*.

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

Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; and Jia, H. 2020. FFA-Net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11908–11915.

Ruan, L.; Bemana, M.; Seidel, H.-p.; Myszkowski, K.; and Chen, B. 2023. Revisiting Image Deblurring with an Efficient ConvNet. *arXiv preprint arXiv:2302.02234*.

Ruan, L.; Chen, B.; Li, J.; and Lam, M. 2022. Learning to Deblur Using Light Field Generated and Real Defocus Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16304–16313.

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

Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32: 1927–1941.

Tsai, F.-J.; Peng, Y.-T.; Lin, Y.-Y.; Tsai, C.-C.; and Lin, C.-W. 2022. Stripformer: Strip Transformer for Fast Image Deblurring. In *Proceedings of the European Conference on Computer Vision*.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. MAXIM: Multi-Axis MLP for Image Processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5769–5780.

Valanarasu, J. M. J.; Yasarla, R.; and Patel, V. M. 2022. TransWeather: Transformer-Based Restoration of Images Degraded by Adverse Weather Conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2353–2363.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, T.; Tao, G.; Lu, W.; Zhang, K.; Luo, W.; Zhang, X.; and Lu, T. 2022a. Restoring Vision in Hazy Weather with Hierarchical Contrastive Learning. *arXiv preprint arXiv:2212.11473*.

Wang, Y.; Li, Y.; Wang, G.; and Liu, X. 2022b. Multi-scale attention network for single image super-resolution. *arXiv preprint arXiv:2209.14145*.

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

Ye, T.; Zhang, Y.; Jiang, M.; Chen, L.; Liu, Y.; Chen, S.; and Chen, E. 2022. Perceiving and Modeling Density for Image Dehazing. In *Proceedings of the European Conference on Computer Vision*, 130–145.

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

Zhang, J.; Cao, Y.; Fang, S.; Kang, Y.; and Wen Chen, C. 2017. Fast haze removal for nighttime image using maximum reflectance prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7418–7426.

Zhang, J.; Cao, Y.; Zha, Z.-J.; and Tao, D. 2020. Nighttime dehazing with a synthetic benchmark. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2355–2363.

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

Zhou, M.; Huang, J.; Guo, C.-L.; and Li, C. 2023. Fourmer: An Efficient Global Modeling Paradigm for Image Restoration. In *International Conference on Machine Learning*, 42589–42601.

Zhu, Z.; Zhang, D.; Wang, Z.; Feng, S.; and Duan, P. 2023. Spectral Dual-Channel Encoding for Image Dehazing. *IEEE Transactions on Circuits and Systems for Video Technology*.

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