Focal Network for Image Restoration

Yuning Cui[1] Wenqi Ren[2*] Xiaochun Cao[2] Alois Knoll[1]
[1]Technical University of Munich [2]Shenzhen Campus of Sun Yat-sen University
{yuning.cui,knoll}@in.tum.de {renwq3,caoxiaochun}@mail.sysu.edu.cn

Abstract

Image restoration aims to reconstruct a sharp image from its degraded counterpart, which plays an important role in many fields. Recently, Transformer models have achieved promising performance on various image restoration tasks. However, their quadratic complexity remains an intractable issue for practical applications. The aim of this study is to develop an efficient and effective framework for image restoration. Inspired by the fact that different regions in a corrupted image always undergo degradations in various degrees, we propose to focus more on the important areas for reconstruction. To this end, we introduce a dual-domain selection mechanism to emphasize crucial information for restoration, such as edge signals and hard regions. In addition, we split high-resolution features to insert multi-scale receptive fields into the network, which improves both efficiency and performance. Finally, the proposed network, dubbed FocalNet, is built by incorporating these designs into a U-shaped backbone. Extensive experiments demonstrate that our model achieves state-of-the-art performance on ten datasets for three tasks, including single-image defocus deblurring, image dehazing, and image desnowing. Our code is available at https://github.com/c-yn/FocalNet.

1. Introduction

Bad weather or physical limitations of cameras will degrade visibility of captured images and further exert a negative impact on robustness of downstream tasks. In this regard, image restoration is immensely useful to remove those undesired degradations, *e.g.,* haze, snowflake, and blur, thus playing an essential role in surveillance, autonomous vehicles, remote sensing, and medical imaging [48]. Due to its ill-posed property, many conventional methods have been proposed by resorting to assumptions and hand-crafted features to reduce the solution space. However, these methods are inapplicable to complicated real-world scenarios [66].

The rapid development of convolutional neural networks (CNNs) has alleviated the above issue to some extent by learning generalizable image priors from large-scale collected datasets [25, 44]. Recently, Transformer models have injected vitality into image restoration and have achieved promising performance by providing powerful ability of modeling long-range pixel interactions, which is useful for recovering large-scale degradations [7, 32, 61]. Despite a few remedies, the quadratic complexity of self-attention still remains a formidable issue for practical applications.

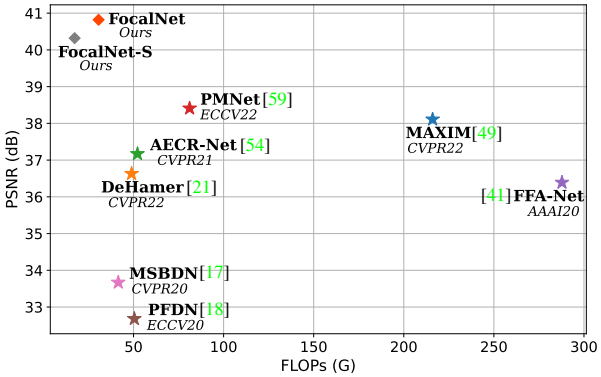

Figure 1. PSNR *vs.* FLOPs on the SOTS-Indoor [27] dataset.

How to effectively capture critical information parsimoniously has long been a key problem in the computer vision and pattern recognition community. Successful examples include attention mechanism [53], focal loss [34], and the recent advance of partial AUC optimization [58]. Inspired by this, in this study, instead of pursuing large receptive field or exploring modification for the Transformer architecture, we aim to develop an efficient and effective CNN-based framework by paying more attention to informative signals for reconstruction, such as edge information or regions that are difficult to recover. In this direction, existing approaches can be roughly divided into two categories: auxiliary training and attention based methods. The former mainly leverages auxiliary techniques or data, *e.g.,* semantic segmentation [12], depth estimation [29], and optical flow estimation [60], to locate degradations or edge information. Nonetheless, these algorithms always need additional com-

*Corresponding Author

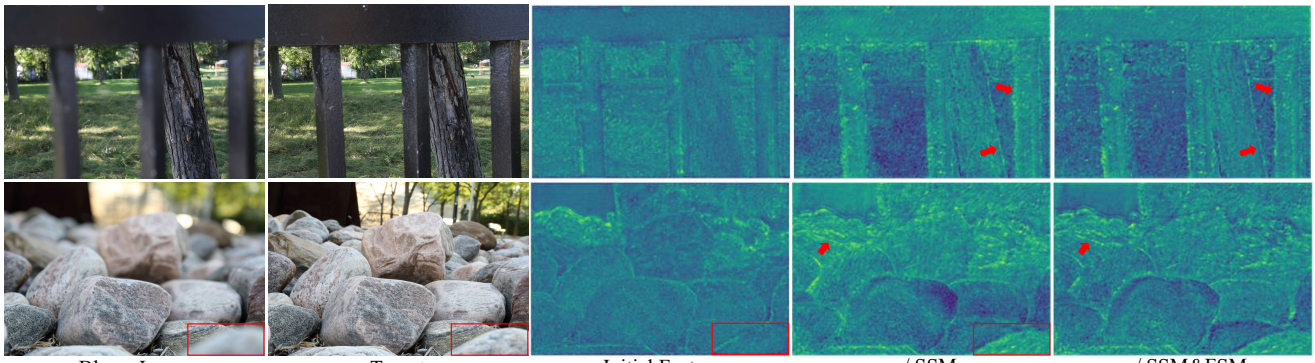

Figure 2. Effects of our dual-domain selection mechanism (DSM). From left to right: blurry images obtained from DPDD [1], ground-truth images, input features of our DSM, results of spatial selection, and results of spatial&frequency selection. SSM helps focus on degradation regions while FSM emphasizes edge information. Zoom in for the best view.

plicated branch and elaborately devised training strategies to generate supervisory information. The other line on this topic is to design attention mechanisms to attend to informative regions or control information transmission [8, 41, 62]. These approaches mostly lie in the spatial domain while ignoring the usage of spectral information, which can also provide useful information for reconstruction.

To prompt models to focus more on critical regions, we propose a novel dual-domain selection mechanism (DSM) by sufficiently leveraging discrepancies between sharp/degraded image pairs in both spatial and spectral domains. Concretely, our mechanism comprises two components: spatial selection module (SSM) and frequency selection module (FSM). SSM takes features as input and determines general locations of degradations for each channel by deploying depthwise convolution layers. Then FSM is used to amplify high-frequency signals or hard regions by removing low frequency from features. The proposed network, FocalNet, is established by incorporating DSM into a U-shaped CNN backbone. To save computation overhead, we only insert DSM into the bottleneck module of our FocalNet, which includes the lowest-resolution features.

Moreover, we split the high-resolution features into two parts over channel dimension. Half features are downsampled to lower resolution, which can not only reduce complexity but also boost performance by providing multi-scale receptive fields for degradations of different sizes.

Based on the above designs, our FocalNet exhibits state-of-the-art performance on three image restoration tasks. For dehazing, FocalNet outperforms PMNet [59] on both synthetic and real-world benchmark datasets with lower computational complexity, as illustrated in Figure 1. For the desnowing task, FocalNet is superior to Transformer-based framework TransWeather [50] on three commonly used desnowing datasets. Our network also shows potential on defocus deblurring problem by producing a performance gain of 0.2 dB PSNR over Restormer [61] on the combined category of DPDD [1] dataset. Overall, the main contributions of this study are summarized as follows:

- We propose a novel dual-domain selection mechanism (DSM) that amplifies the response of important regions to assist in recovering clean features.
- We develop an efficient and effective focal network that provides multi-scale representation learning for image restoration.
- Extensive experiments on ten datasets demonstrate that the proposed network, FocalNet, performs favorably against state-of-the-art algorithms on three representative image restoration tasks.

2. Related Work

Image Restoration Architectures. As a long-standing task, image restoration aims to remove undesired degradations in corrupted images, which plays an important role in many fields, such as robot vision, medical applications, and surveillance [48, 66]. Recently, CNN-based architectures have significantly advanced the performance compared to conventional methods [1, 8, 14, 16]. Among these architectures, the encoder-decoder paradigm is a popular solution to learn hierarchical representations [39, 44]. In addition, a great number of functional units have been developed or borrowed from other realms, such as dilated convolution [68], skip connections [13], dynamic filter [25], and various attention mechanisms [41]. More recently, Transformer models have been imported into low-level vision tasks and provided promising performance [7, 32]. Thereafter, a few measures have been taken to reduce the computational complexity of self-attention by restricting the operation region [32] or switching the operation dimension [61].

Spectral Networks. Apart from spatial representation learning, numerous deep frameworks have been proposed

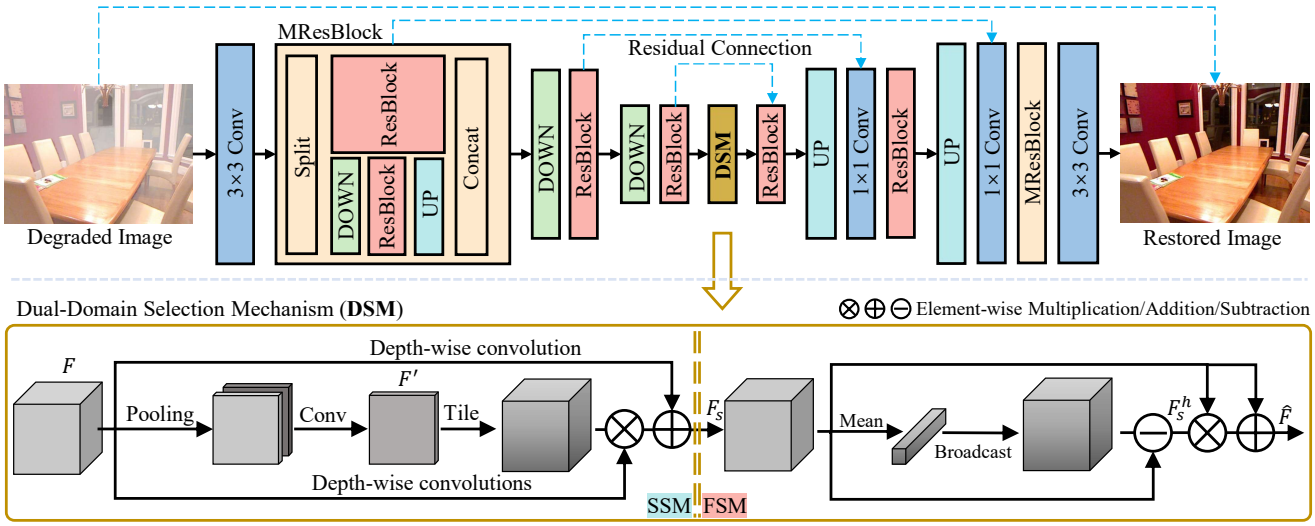

Figure 3. Architecture of the proposed FocalNet with dual-domain selection mechanism (DSM) that consists of two components, *i.e.,* spatial selection module (SSM) and frequency selection module (FSM). ResBlock contains n residual blocks, comprising two 3×3 convolution layers and activation function in between.

to bridge frequency gaps between sharp/degraded image pairs [15, 39, 68]. The common practice is to decompose features into different frequency components via transformation tools such as wavelet transform [10, 68], Fourier transform [22, 39], pooling technique [15], and conventional filters [35], and then treat each component individually via convolution layers. In addition, a few researches have investigated the different roles of phase and amplitude, and proposed architectures to recover them separately [28]. In our work, we simply remove the lowest-frequency signal from the resulting features of SSM to provide guidance for further reconstruction.

Auxiliary Training. In addition to the provided ground-truth images in training sets of image restoration tasks, a great number of networks have been proposed to resort to auxiliary supervision. Semantic priors have been introduced into low-level tasks to provide color, boundary or location information [12, 43, 45]. However, global semantic priors are less effective for degradations caused by large depth variations. Thus many methods have proposed to estimate depth map to generate edge and structure signals for restoration [24, 29]. Moreover, there are many works that integrate other auxiliary information such as optical flow [56, 60] and event data [57]. However, the above solutions always require additional data, expensive convolution branch, and complicated training strategies [45].

3. Method

In this section, we first delineate the overall architecture of FocalNet. Then we describe our modules: Multi-scale ResBlock (MResBlock) and Dual-domain Selection Mechanism (DSM). Finally, we detail the loss functions.

3.1. Overall Pipeline

As illustrated in Figure 3, the proposed FocalNet adopts the popular encoder-decoder architecture to learn hierarchical representations efficiently. Both encoder and decoder networks consist of three scales. In our paper, we refer to the first scale as the sub-network that involves the highest-resolution features. MResBlock constitutes the main part of the first scale. The other two scales are mainly composed of ResBlock, which consists of n residual blocks. Given a degraded image of size $H \times W \times 3$, where $H \times W$ and C represent spatial locations and the number of channels respectively, a 3×3 convolution layer is used to extract shallow features of size $H \times W \times C$. Then, the shallow features pass through three-scale symmetric encoder-decoder and are transformed into restored features, *i.e.,* output features of MResBlock in the first scale of decoder. Starting from the highest-resolution input, the encoder gradually reduces spatial size and expands the number of channels. The decoder does the opposite to restore clean features from the deepest features. During this process, the decoder features are concatenated with the encoder features to assist recovery, followed by a 1×1 convolution to adjust channel dimension. Finally, the predicted clean image is generated by the last 3×3 convolution layer and image-level residual connection. Upsampling (UP) and downsampling (DOWN) operations are implemented by transposed and strided convolutions except for the upsampling layer in MResBlock that adopts *bilinear* interpolation. The proposed DSM is injected into the bottleneck location to select the most im-

portant regions for reconstruction. In addition, we apply the multi-input and multi-output strategies to ease training difficulty following previous methods [13, 15, 39, 49].

3.2. Multi-scale ResBlock (MResBlock)

Recently, pursuing multi-scale receptive fields is a hot topic in computer vision community [5, 20, 42, 55], especially for Transformer-based models [33, 42]. Inspired by [11, 40, 46], we adopt multi-scale mechanism in ResBlock to form our MResBlock via splitting and downsampling operations, as illustrated in Figure 3. Specifically, given input features, we first split them equally along channel dimension into two components. Next, half features are reduced to a quarter of the original resolution using strided convolution. The resulting features are fed into ResBlock, and then up-sampled to the original size. The other half are directly processed by ResBlock. The final output of MResBlock is obtained by concatenating resulting features of two branches. MResBlock enjoys two main advantages. Firstly, it improves performance by realizing multi-scale representation learning for degradations of different sizes and enhances spectral learning for different frequencies [40]. Secondly, it improves efficiency by reducing feature resolution.

3.3. Dual-domain Selection Mechanism (DSM)

The main goal of this study is to develop an efficient network for image restoration by focusing on the more important regions. This aim is achieved by the proposed DSM, which amplifies the response of informative information in two domains (see Figure 2). As illustrated in the bottom part of Figure 3, it consists of two components: spatial selection module (SSM) and frequency selection module (FSM). Given input features $F \in \mathbb{R}^{H\times W\times C}$, SSM and FSM are employed successively, which can be expressed as:

$$\hat{F} = \mathrm{FSM}(\mathrm{SSM}(F)). \tag{1}$$

Next, we introduce these two elements in details.

Spatial Selection Module (SSM). SSM helps the network focus on important regions in the spatial domain, providing initial locations of severe degradations for subsequent FSM. Our SSM has three branches. The main path is built upon the CBAM [53] to produce the general feature representation for locations of degradations to focus. Specifically, given an intermediate feature map F, we first squeeze F along channel dimension through two types of pooling techniques, *i.e.,* max pooling and average pooling, and then generate the general feature map via a convolution layer, which is formally expressed as:

$$F' = \mathrm{Conv}_3([\mathrm{AvgPool}(F), \mathrm{MaxPool}(F)]), \tag{2}$$

where $[\cdot,\cdot]$ indicates concatenation; AvgPool, MaxPool and Conv_3 represent average pooling, max pooling, and convolution layer of 3×3 kernel size. By doing this, $F' \in \mathbb{R}^{H\times W\times 1}$ contains degradation locations to focus [53].

Since each channel differs in degradation patterns, we further generate channel-wise representation by performing the channel-separated transformation for the input feature F via depth-wise convolutions, and then modulate resulting features with F'. This process is expressed as follows:

$$F_s = \mathrm{DConvs}_{5,7}(F) \otimes \mathrm{T}(F', C) + \mathrm{DConv}_3(F), \tag{3}$$

where $\mathrm{DConvs}_{5,7}$ denotes cascaded depth-wise convolution layers of kernel sizes 5×5 and 7×7; DConv_3 represents depth-wise convolution with 3×3 kernel; $\otimes$ indicates an element-wise multiplication; and $\mathrm{T}(F', C)$ is the tile function that copies F' for C times along the channel dimension to $\mathbb{R}^{H\times W\times C}$. We then feed the spatially selected features $F_s \in \mathbb{R}^{H\times W\times C}$ to FSM for frequency selection.

Frequency Selection Module (FSM). We can directly utilize F_s to assist the recovery process. Motivated by the fact that degraded/sharp image pairs have similar low-frequency components, while differing at high frequencies, we further emphasize regions that contain the real difference between input/sharp image pairs by removing lowest frequency via the proposed FSM. To this end, we first apply the mean filter to F_s to generate low-frequency features, and then obtain the complementary high-frequency features by subtracting the resulting low-frequency signals from the input, which is expressed by:

$$F_s^h = F_s - \mathrm{Mean}(F_s). \tag{4}$$

In our case, the mean filter is implemented by the channel-wise global average pooling. The final output of FSM/DSM is generated using the element-wise multiplication between F_s^h and F_s, and residual connection, which is expressed as:

$$\hat{F} = F_s^h \otimes F_s + F_s. \tag{5}$$

After DSM, the important regions are emphasized, *e.g.,* edge signals in Figure 2 for defocus deblurring.

3.4. Loss Function

To facilitate the selection process in both spatial and frequency domains, we adopt dual-domain l_1 loss functions following [13, 15]. For each output/target image pair with the same resolution, loss functions are given by:

$$\mathcal{L}_s = \frac{1}{P}\|\hat{I} - G\|_1 \tag{6}$$

$$\mathcal{L}_f = \frac{1}{P}\|\mathcal{F}(\hat{I}) - \mathcal{F}(G)\|_1 \tag{7}$$

$$\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_f \tag{8}$$

where $\hat{I}$ and G denote the output and ground-truth images, respectively; P indicates total elements for normalization; $\mathcal{F}$ represents fast Fourier transform; and λ is empirically set to 0.1 for balancing dual-domain training.

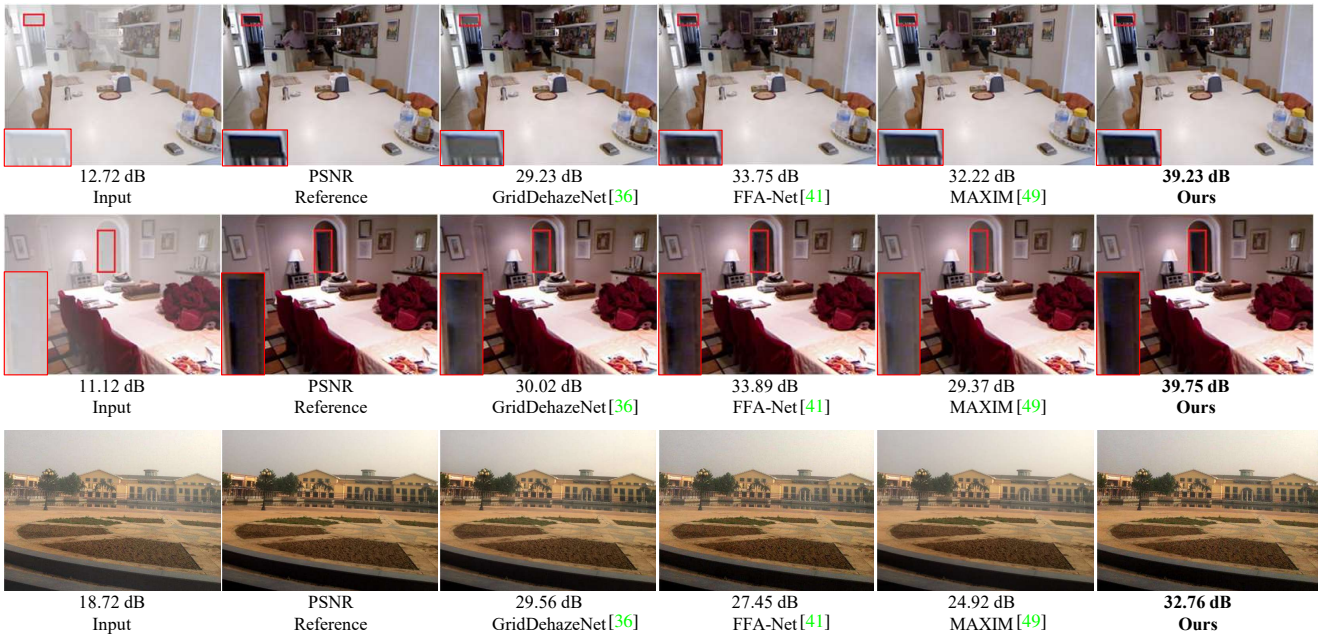

Figure 4. Image dehazing comparisons on the SOTS [27] test sets.

4. Experiments and Analysis

We evaluate FocalNet on 10 datasets for three image restoration tasks, including image dehazing, image desnowing, and single-image defocus deblurring. We provide experimental results for image motion deblurring and image denoising in the supplementary material.

4.1. Datasets and Evaluation Protocol

We measure the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity index (SSIM) [52] between predicted results and ground-truth images for all datasets. Mean Absolute Error (MAE) and Learned Perceptual Image Patch Similarity (LPIPS) [67] are additionally adopted for defocus deblurring. FLOPs are measured on 256×256 patch. In tables, the best performance is marked in **bold**.

Image Dehazing. We train and evaluate our models on both synthetic and real-world datasets for image dehazing. Following [21, 59], we train separate models on the RESIDE-Indoor and RESIDE-Outdoor datasets [27], and evaluate the resulting models on the corresponding test sets of RESIDE, *i.e.,* SOTS-Indoor and SOTS-Outdoor, respectively. In addition, we adopt three real-world datasets, *i.e.,* Dense-Haze [2], NH-HAZE [3], and O-HAZE [4], to verify the robustness of our model in more challenging real-world scenarios. Apart from the above daytime dehazing datasets, we also demonstrate the effectiveness of FocalNet on the nighttime dataset NHR [65].

Image Defocus Deblurring. Consistent with previous algorithms [25, 44, 61], we use the DPDD [1] dataset that consists of 350 scenes for training, 74 scenes for validation, and 76 scenes for testing. Each scene comprises four images, labeled as center view, left view, right view, and all-in-focus ground-truth images. FocalNet is trained by taking the center view image as input and calculating loss values between the predicted clean image and ground truth.

Image Desnowing. For the desnowing problem, we train and evaluate our models on three commonly used datasets, *i.e.,* SRRS [9], CSD [10], and Snow100K [37]. Dataset selection and testing methods are consistent with the recent algorithm [9] for a fair and convincing comparison.

4.2. Implementation Details

Unless stated otherwise, the following hyper-parameters are adopted. Depending on the task complexity, we scale the model by varying the number of residual blocks in all ResBlock, *i.e.,* 16 for deblurring and 4 for dehazing/desnowing. In our FocalNet, we only deploy MResBlock in the first scale of encoder/decoder. Convolution parameters are not shared between two branches of MResBlock. The models are trained using Adam [23] with initial learning rate as $8e^{-4}$, which is gradually reduced to $1e^{-6}$ with cosine annealing [38]. For data augmentation, we adopt random horizontal flips with a probability of 0.5. Models are trained on 32 samples of size 256×256 for each iteration. More details of training configuration for each specific dataset are provided in the supplementary material.

4.3. Image Dehazing Results

Quantitative Comparisons. We report the quantitative performance of image dehazing approaches on both syn-

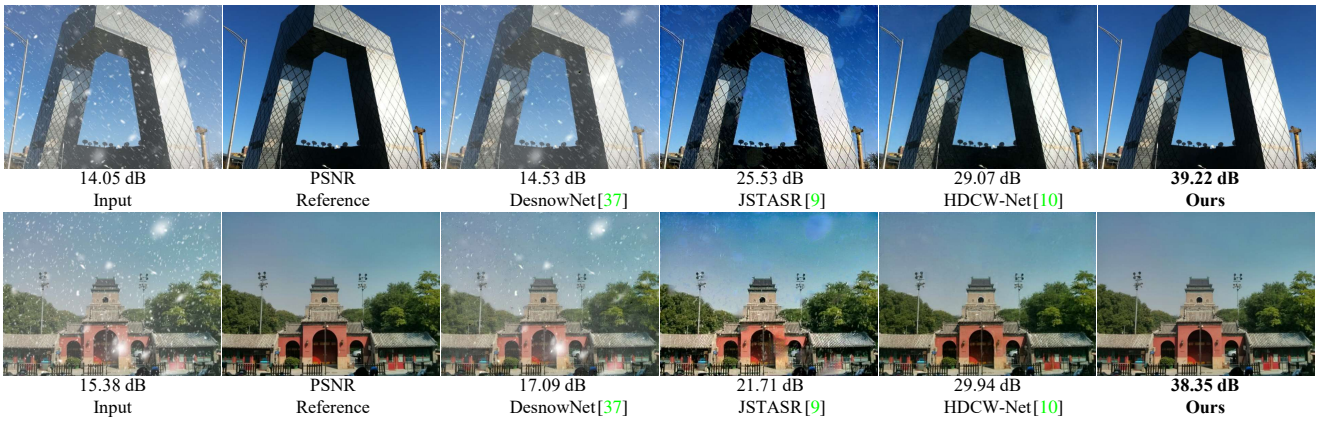

Figure 5. Image desnowing comparisons on the CSD [10] dataset. Our method is more effective in snow removal than other algorithms.

Method	SOTS-Indoor [27]		SOTS-Outdoor [27]		Dense-Haze [2]		NH-HAZE [3]		O-HAZE [4]		Params (M)	FLOPs (G)
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
DehazeNet [6]	19.82	0.821	24.75	0.927	13.84	0.43	16.62	0.52	17.57	0.77	0.009	0.581
AOD-Net [26]	20.51	0.816	24.14	0.920	13.14	0.41	15.40	0.57	15.03	0.54	0.002	0.115
GridDehazeNet [36]	32.16	0.984	30.86	0.982	-	-	13.80	0.54	-	-	0.956	21.49
MSBDN [17]	33.67	0.985	33.48	0.982	15.37	0.49	19.23	0.71	24.36	0.75	31.35	41.54
FFA-Net [41]	36.39	0.989	33.57	0.984	14.39	0.45	19.87	0.69	22.12	0.77	4.456	287.8
AECR-Net [54]	37.17	0.990	-	-	15.80	0.47	19.88	0.72	-	-	2.611	52.20
DeHamer [21]	36.63	0.988	35.18	0.986	16.62	0.56	**20.66**	0.68	-	-	132.45	48.93
PMNet[59]	38.41	0.990	34.74	0.985	16.79	0.51	20.42	0.73	24.64	0.83	18.90	81.13
FocalNet (Ours)	**40.82**	**0.996**	**37.71**	**0.995**	**17.07**	**0.63**	20.43	**0.79**	**25.50**	**0.94**	3.74	30.63

Table 1. Image dehazing results on both synthetic dataset [27] and real-world datasets [2, 3, 4].

Method	NDIM [64]	GS [31]	MRPF [63]	MRP [63]	OSFD [65]	HCD [51]	FocalNet Ours
PSNR	14.31	17.32	16.95	19.93	21.32	23.43	**25.35**
SSIM	0.526	0.629	0.667	0.777	0.804	0.953	**0.969**

Table 2. Nighttime image dehazing results on NHR [65] dataset.

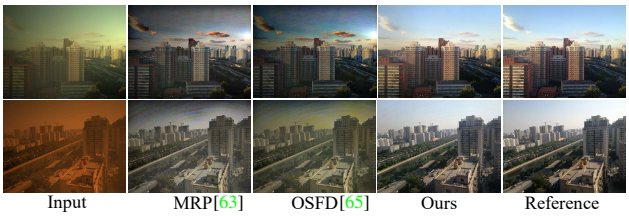

Figure 6. Nighttime dehazing comparisons on NHR [65] dataset.

thetic [27] and real-world [2, 3, 4] datasets in Table 1. Overall, our method receives better performance on all datasets than other state-of-the-art algorithms. Specifically, on the daytime synthetic dataset SOTS-Indoor [27], our method outperforms PMNet [59] by 2.41 dB PSNR with only 20% parameters and 38% FLOPs. Furthermore, our model yields a significant performance gain of 2.53 dB in terms of PSNR over Transformer model DeHamer [21] on SOTS-Outdoor [27] with 97% fewer parameters.

In addition, our method is well generalized to the more challenging real-world scenarios and obtains the best performance on most metrics. Particularly on the O-HAZE [4] dataset, FocalNet yields considerable gain of 0.86 dB PSNR and 0.11 SSIM over PMNet [59].

Apart from daytime dehazing datasets, we also evaluate the effectiveness of our model on the nighttime dehazing dataset NHR [65]. As shown in Table 2, FocalNet significantly outperforms the recent HCD [51] by 1.92 dB PSNR. The results demonstrate the efficacy of our design.

Method	CSD [10]		SRRS [9]		Snow100K [37]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DesnowNet [37]	20.13	0.81	20.38	0.84	30.50	0.94
CycleGAN [19]	20.98	0.80	20.21	0.74	26.81	0.89
All in One [30]	26.31	0.87	24.98	0.88	26.07	0.88
JSTASR [9]	27.96	0.88	25.82	0.89	23.12	0.86
HDCW-Net [10]	29.06	0.91	27.78	0.92	31.54	**0.95**
TransWeather [50]	31.76	0.93	28.29	0.92	31.82	0.93
NAFNet [8]	33.13	0.96	29.72	0.94	32.41	**0.95**
FocalNet (Ours)	**37.18**	**0.99**	**31.34**	**0.98**	**33.53**	**0.95**

Table 3. Image desnowing results on three widely used datasets.

Visual Comparisons. The daytime and nighttime visual results produced by several dehazing methods are illustrated

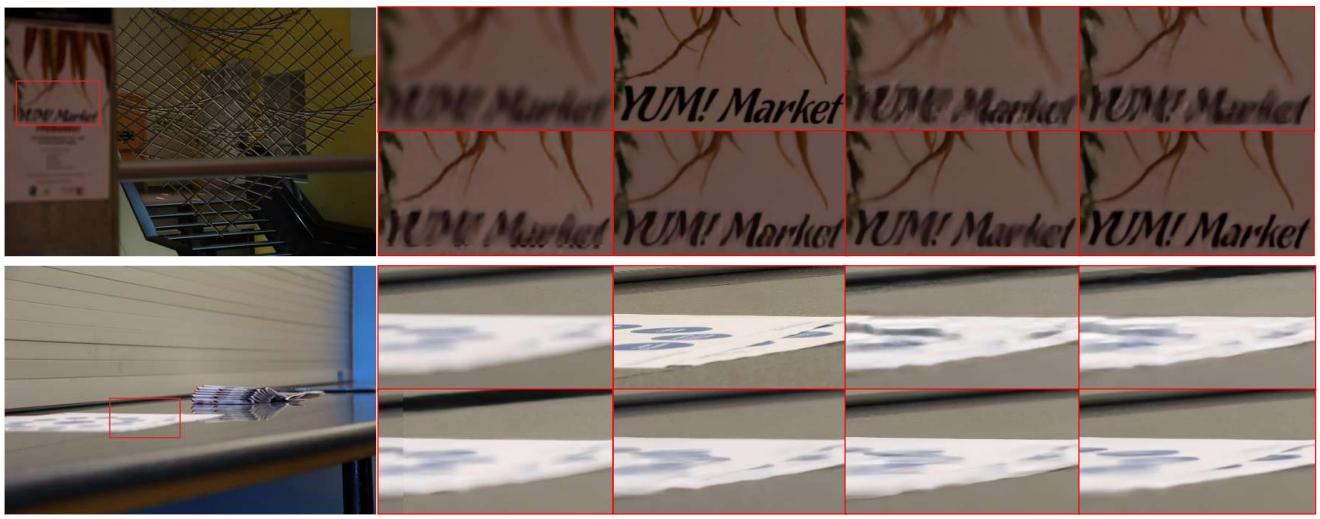

Figure 7. Single-image defocus deblurring results on the DPDD [1] dataset. We display the magnified parts of the produced images for clarity. From left-top to right-bottom: input blurry images, ground-truth images, and the predicted images obtained by KPAC [47], IFAN [25], DeepRFT [39], DRBNet [44], Restormer [61], and our FocalNet, respectively.

	Indoor Scenes				Outdoor Scenes				Combined			
Method	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓
DPDNet [1]	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 [47]	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227
DeepRFT [39]			-				-		25.71	0.801	0.039	0.218
IFAN [25]	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217
DRBNet [44]			-				-		25.73	0.791	-	0.183
Restormer [61]	28.87	**0.882**	0.025	**0.145**	23.24	**0.743**	0.050	**0.209**	25.98	**0.811**	0.038	**0.178**
FocalNet (Ours)	**29.10**	0.876	**0.024**	0.173	**23.41**	**0.743**	**0.049**	0.246	**26.18**	0.808	**0.037**	0.210

Table 4. Single-image defocus deblurring results on the DPDD [1] dataset.

in Figure 4 and Figure 6, respectively. Our method is more effective in removing haze blurs in both daytime (indoor and outdoor) and nighttime scenes than other algorithms, such as blurs on the doors in top two images of Figure 4.

4.4. Image Desnowing Results

Quantitative Comparisons. The numerical results on three desnowing datasets, *i.e.,* SRRS [9], CSD [10], and Snow100K [37], are reported in Table 3. As can be seen, our model shows strong ability on snow removal by consistently achieving better or comparable PSNR/SSIM scores on all metrics. Compared to NAFNet [8], FocalNet obtains significant performance gains of 1.62 dB and 4.05 dB PSNR on the SRRS [9] and lately proposed CSD [10] datasets, respectively. In addition, our model yields much higher scores than TransWeather [50] with 17% parameters.

Visual Comparisons. The visual results are shown in Figure 5. Images produced by our model are of high quality without artifacts and visually closer to ground-truth images.

4.5. Image Defocus Deblurring Results

Quantitative Comparisons. The quantitative comparisons on the DPDD [1] dataset are provided in Table 4. Our method performs favorably against state-of-the-art algorithms. Specifically, FocalNet obtains significant gains over the CNN-based approaches on the combined category, *i.e.,* 0.45 dB over DRBNet [44] and 0.47 dB over DeepRFT [39]. In addition, compared to Transformer model Restormer [61], our method still obtains a performance improvement of 0.2 dB PSNR with only half parameters.

Visual Comparisons. Figure 7 illustrates visual results. Compared to other approaches, our model recovers more fine details, such as patterns on pages.

4.6. Ablation Studies

We conduct ablation studies to demonstrate the effectiveness of our modules by training the tiny model on RESIDE-Indoor [27] and testing on SOTS-Indoor [27]. The number of residual block is set to 1 in all ResBlock. The model is

	MResBlock	SSM	FSM	PSNR	FLOPs/G	Params/M
(a)				31.33	15.44	1.48
(b)		✓		33.51	15.48	1.49
(c)			✓	32.44	15.44	1.49
(d)		✓	✓	35.56	15.48	1.49
(e)	✓	✓		33.71	13.82	1.47
(f)	✓		✓	32.60	13.77	1.47
(g)	✓	✓	✓	**35.60**	13.82	1.47

Table 5. Ablation studies for different components of FocalNet on the SOTS-Indoor [27] dataset.

	Method	PSNR	FLOPs/G	Params/M
(a)	Simple Gate [8]	32.23	13.81	1.47
(b)	Supervised Attention [62]	32.29	14.98	1.76
(c)	CBAM [53]	32.71	13.77	1.46
(d)	SSM	**33.71**	13.82	1.47
(e)	CBAM [53]+FSM	33.23	13.77	1.46
(f)	SSM+FSM	**35.60**	13.82	1.47

Table 6. Comparisons with other attention mechanisms.

trained for only 300 epochs with the initial learning rate as $1e^{-4}$ and batch size as 4. Other settings are identical with that of our final dehazing model. The baseline network is obtained by substituting ResBlock for MResBlock and removing DSM from the tiny model. More ablation studies are provided in the supplementary material.

Effects of Individual Components. As shown in Table 5a, the baseline receives 31.33 dB PSNR. SSM (Table 5b) and FSM (Table 5c) yield accuracy gains of 2.18 dB and 1.11 dB over the baseline, respectively. Equipped with DSM, the model (Table 5d) receives further performance boost by 4.23 dB over the baseline. By replacing ResBlock with MResBlock in the first scale, models consistently receive higher scores than those without MResBlock. Specifically, our choice (Table 5g) outperforms the counterpart (Table 5d) by 0.04 dB with lower computation overhead.

In addition, the visual results of our DSM are illustrated in Figure 2 (defocus debluring) and Figure 8 (dehazing). As shown in Figure 2, SSM helps the model focus more on the severe degradation regions, *e.g.,* metal fence. FSM further highlights the edge signals by removing low-frequency information. For dehazing, the hard regions that are difficult to recover are emphasized by our DSM (see Figure 8). More examples are available in the supplementary material.

Comparisons with Alternatives to SSM. We further demonstrate superiority of our SSM by replacing it with three popular attention mechanisms. As represented in Table 6, the simple gate [8] and supervised attention [62] lead to performance degradation from 33.71 to 32.23 dB and 32.29 dB PSNR, respectively. Compared to CBAM [53], which includes both spatial and channel attention, our SSM

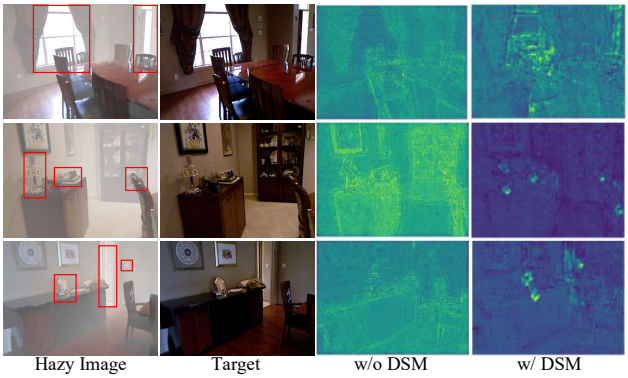

Figure 8. Visual results of DSM for dehazing. From left to right: hazy images obtained from SOTS-Indoor [27], ground-truth images, input and output features of DSM. The complicated regions worthy of more attention are highlighted by our DSM.

Method	PSNR	FLOPs/G	Params/M
(a) FocalNet w/o MResBlock	40.69	33.96	3.76
(b) FocalNet	**40.82**	30.63	3.74
(c) FocalNet-S	40.32	17.33	2.27

Table 7. Design choices for MResBlock.

(Table 6d) receives 1 dB performance gain with comparable computation overhead. The similar conclusion can also be drawn by comparing models of Table 6e and Table 6f, where FSM is additionally used. It is worth mentioning that our FSM boosts the performance when cooperating with CBAM (see Table 6c and Table 6e), demonstrating the efficacy and generalization ability of our design.

Order of FSM and SSM. When we swap the order in which FSM and SSM are used, the performance drops from 35.60 dB (Table 5g) to 35.17 dB PSNR. This phenomenon verifies the efficacy of our design, where we first apply SSM to attend to general degradation regions and then leverage FSM to emphasize the more important parts, such as edge signal in Figure 2 and hard regions in Figure 8.

Design Choices for MResBlock. We study the effect of the multi-scale mechanism in our final dehazing model. The training settings are identical with FocalNet in Table 1. As represented in Table 7, MResBlock leads to performance gain of 0.13 dB PSNR with high efficiency. We further explore the potential of MResBlock by using it in all scales of encoder/decoder to form FocalNet-S (Figure 7c). This version degrades the performance by 0.5 dB compared to FocalNet, which is probably because the disadvantage of losing spatial information caused by reducing size of low-resolution features outweighs the advantage of multi-scale learning. It is worth mentioning that FocalNet-S still achieves 40.32 dB PSNR with much lower computation overhead compared to other algorithms (see Figure 1).

5. Conclusion

In this study, we present a focal network for image restoration, dubbed FocalNet, which is effective and computationally efficient. The core idea of our work is to focus on the important regions for reconstruction. To this end, we propose two modules: SSM and FSM. SSM is built on spatial attention to detect the degradation regions for subsequent frequency selection. FSM further emphasizes the edge signals or regions that are difficult to recover. By deploying two modules successively, the network is capable of paying more attention to regions that really matter to reconstruction. In addition, we insert the multi-scale mechanism into the network by reducing resolution of half the channels of input feature. This design not only improves performance but also reduces complexity. Experiments on 10 datasets demonstrate that our model achieves state-of-the-art performance for several image restoration tasks.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 62025604, U1803264) and Shenzhen Science and Technology Program (No. 20220016).

References

[1] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pages 111–126, 2020. 2, 5, 7

[2] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE International Conference on Image Processing*, pages 1014–1018, 2019. 5, 6

[3] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte. Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 5, 6

[4] Codruta O. Ancuti, Cosmin Ancuti, Radu Timofte, and Christophe De Vleeschouwer. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 5, 6

[5] Shaojie Bai, Vladlen Koltun, and J Zico Kolter. Multiscale deep equilibrium models. volume 33, pages 5238–5250, 2020. 4

[6] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 6

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 1, 2

[8] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33, 2022. 2, 6, 7, 8

[9] Wei-Ting Chen, Hao-Yu Fang, Jian-Jiun Ding, Cheng-Che Tsai, and Sy-Yen Kuo. Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In *European Conference on Computer Vision*, pages 754–770, 2020. 5, 6, 7

[10] Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I Chen, Jian-Jiun Ding, Sy-Yen Kuo, et al. All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4196–4205, 2021. 3, 5, 6, 7

[11] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 4

[12] Ziang Cheng, Shaodi You, Viorela Ila, and Hongdong Li. Semantic single-image dehazing. *arXiv preprint arXiv:1804.05624*, 2018. 1, 3

[13] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4641–4650, 2021. 2, 4

[14] Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. Irnext: Rethinking convolutional network design for image restoration. In *International Conference on Machine Learning*, 2023. 2

[15] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *International Conference on Learning Representations*, 2023. 3, 4

[16] Yuning Cui, Yi Tao, Wenqi Ren, and Alois Knoll. Dual-domain attention for image deblurring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 479–487, 2023. 2

[17] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 6

[18] Jiangxin Dong and Jinshan Pan. Physics-based feature dehazing networks. In *European Conference on Computer Vision*, pages 188–204, 2020. 1

[19] Deniz Engin, Anil Genc, and Hazim Kemal Ekenel. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 6

[20] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new

multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662, 2019. 4

[21] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5812–5820, 2022. 1, 5, 6

[22] Shi Guo, Hongwei Yong, Xindong Zhang, Jianqi Ma, and Lei Zhang. Spatial-frequency attention for image denoising. *arXiv preprint arXiv:2302.13598*, 2023. 3

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[24] Dongwoo Lee, Haesol Park, In Kyu Park, and Kyoung Mu Lee. Joint blind motion deblurring and depth estimation of light field. In *Proceedings of the European Conference on Computer Vision*, 2018. 3

[25] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2034–2042, 2021. 1, 2, 5, 7

[26] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 6

[27] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing*, 28(1):492–505, 2018. 1, 5, 6, 7, 8

[28] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. *arXiv preprint arXiv:2302.11831*, 2023. 3

[29] Lerenhan Li, Jinshan Pan, Wei-Sheng Lai, Changxin Gao, Nong Sang, and Ming-Hsuan Yang. Dynamic scene deblurring by depth guided model. *IEEE Transactions on Image Processing*, 29:5273–5288, 2020. 1, 3

[30] Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in one bad weather removal using architectural search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 6

[31] Yu Li, Robby T Tan, and Michael S Brown. Nighttime haze removal with glow and multiple light colors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 226–234, 2015. 6

[32] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1833–1844, 2021. 1, 2

[33] Yuxuan Liang, Pan Zhou, Roger Zimmermann, and Shuicheng Yan. Dualformer: Local-global stratified transformer for efficient video recognition. In *European Conference on Computer Vision*, pages 577–595, 2022. 4

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 1

[35] Keng-Hao Liu, Chia-Hung Yeh, Juh-Wei Chung, and Chuan-Yu Chang. A motion deblur method based on multi-scale high frequency residual image learning. *IEEE Access*, 8:66025–66036, 2020. 3

[36] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7314–7323, 2019. 5, 6

[37] Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. Desnownet: Context-aware deep network for snow removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, 2018. 5, 6, 7

[38] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[39] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021. 2, 3, 4, 7

[40] Zizheng Pan, Jianfei Cai, and Bohan Zhuang. Fast vision transformers with hilo attention. In *Advances in Neural Information Processing Systems*, 2022. 4

[41] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11908–11915, 2020. 1, 2, 5, 6

[42] Sucheng Ren, Daquan Zhou, Shengfeng He, Jiashi Feng, and Xinchao Wang. Shunted self-attention via multi-scale token aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10853–10862, 2022. 4

[43] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *IEEE Transactions on Image Processing*, 28(4):1895–1908, 2018. 3

[44] Lingyan Ruan, Bin Chen, Jizhou Li, and Miuling Lam. Learning to deblur using light field generated and real defocus images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16304–16313, 2022. 1, 2, 5, 7

[45] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Deep semantic face deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 3

[46] Chenyang Si, Weihao Yu, Pan Zhou, Yichen Zhou, Xinchao Wang, and Shuicheng YAN. Inception transformer. In *Advances in Neural Information Processing Systems*, 2022. 4

[47] Hyeongseok Son, Junyong Lee, Sunghyun Cho, and Seungyong Lee. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2642–2650, 2021. 7

[48] Jingwen Su, Boyan Xu, and Hujun Yin. A survey of deep learning approaches to image restoration. *Neurocomputing*, 487:46–65, 2022. 1, 2

[49] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022. 1, 4, 5

[50] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022. 2, 6, 7

[51] Tao Wang, Guangpin Tao, Wanglong Lu, Kaihao Zhang, Wenhan Luo, Xiaoqin Zhang, and Tong Lu. Restoring vision in hazy weather with hierarchical contrastive learning. *arXiv preprint arXiv:2212.11473*, 2022. 6

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[53] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *European Conference on Computer Vision*, 2018. 1, 4, 8

[54] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021. 1, 6

[55] Yu-Huan Wu, Yun Liu, Xin Zhan, and Ming-Ming Cheng. P2t: Pyramid pooling transformer for scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[56] Yanyang Yan, Qingbo Wu, Bo Xu, Jingang Zhang, and Wenqi Ren. Vdflow: Joint learning for optical flow and video deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 3

[57] Dan Yang and Mehmet Yamac. Motion aware double attention network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1113–1123, 2022. 3

[58] Zhiyong Yang, Qianqian Xu, Shilong Bao, Yuan He, Xiaochun Cao, and Qingming Huang. Optimizing two-way partial auc with an end-to-end framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[59] Tian Ye, Yunchen Zhang, Mingchao Jiang, Liang Chen, Yun Liu, Sixiang Chen, and Erkang Chen. Perceiving and modeling density for image dehazing. In *European Conference on Computer Vision*, pages 130–145, 2022. 1, 2, 5, 6

[60] Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3

[61] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 1, 2, 5, 7

[62] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 2, 8

[63] Jing Zhang, Yang Cao, Shuai Fang, Yu Kang, and Chang Wen Chen. Fast haze removal for nighttime image using maximum reflectance prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6

[64] Jing Zhang, Yang Cao, and Zengfu Wang. Nighttime haze removal based on a new imaging model. In *IEEE International Conference on Image Processing*, pages 4557–4561, 2014. 6

[65] Jing Zhang, Yang Cao, Zheng-Jun Zha, and Dacheng Tao. Nighttime dehazing with a synthetic benchmark. In *Proceedings of the ACM International Conference on Multimedia*, pages 2355–2363, 2020. 5, 6

[66] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022. 1, 2

[67] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 5

[68] Wenbin Zou, Mingchao Jiang, Yunchen Zhang, Liang Chen, Zhiyong Lu, and Yi Wu. Sdwnet: A straight dilated network with wavelet transformation for image deblurring. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1895–1904, 2021. 2, 3