Strip Attention for Image Restoration

Yuning Cui[1] , **Yi Tao**[2] , **Luoxi Jing**[3] and **Alois Knoll**[1]
[1]School of Computation, Information and Technology, Technical University of Munich, Germany
[2]MIT Universal Village Program, USA
[3]School of Computer Science, Peking University, China
{yuning.cui, knoll}@in.tum.de, yitao@universal-village.org, jingluoxi@stu.pku.edu.cn

Abstract

As a long-standing task, image restoration aims to recover the latent sharp image from its degraded counterpart. In recent years, owing to the strong ability of self-attention in capturing long-range dependencies, Transformer based methods have achieved promising performance on multifarious image restoration tasks. However, the canonical self-attention leads to quadratic complexity with respect to input size, hindering its further applications in image restoration. In this paper, we propose a Strip Attention Network (SANet) for image restoration to integrate information in a more efficient and effective manner. Specifically, a strip attention unit is proposed to harvest the contextual information for each pixel from its adjacent pixels in the same row or column. By employing this operation in different directions, each location can perceive information from an expanded region. Furthermore, we apply various receptive fields in different feature groups to enhance representation learning. Incorporating these designs into a U-shaped backbone, our SANet performs favorably against state-of-the-art algorithms on several image restoration tasks. The code is available at https://github.com/c-yn/SANet.

1 Introduction

Image restoration aims to reconstruct a high-quality image from the observation suffering from various degradations (*e.g.*, blur, snowflake, haze), playing an essential role in many fields, such as surveillance, medical imaging, and remote sensing. It is an inverse problem and has an ill-posed nature. To resolve this challenging problem, a multitude of conventional algorithms have been developed based on hand-crafted features, which are impractical in more complicated real-world scenarios [Zhang *et al.*, 2022].

In recent years, convolutional neural networks (CNNs) have witnessed a significant development of image restoration and achieved remarkable performance compared to traditional approaches by virtue of the powerful mapping capability. A great number of CNN-based methods have been proposed for varied image restoration tasks by designing or borrowing advanced units, including U-shaped backbone [Lee *et al.*, 2021], residual connection [Cho *et al.*, 2021], dilated convolution [Son *et al.*, 2021], and attention modules [Qin *et al.*, 2020; Cui *et al.*, 2023b]. Nevertheless, CNN has two defects that are not beneficial for image restoration: **(a)** The convolution operator has static filters that are not applicable to the dynamic and non-uniform blur. **(b)** The convolution filter has a limited receptive field that is not capable of modeling long-range pixels interactions for large-size blur. Despite many efforts to enlarge the receptive field by stacking deep layers or using dilated convolution [Son *et al.*, 2021], these remedies entail heavy computation burden and still struggle to obtain the global receptive field.

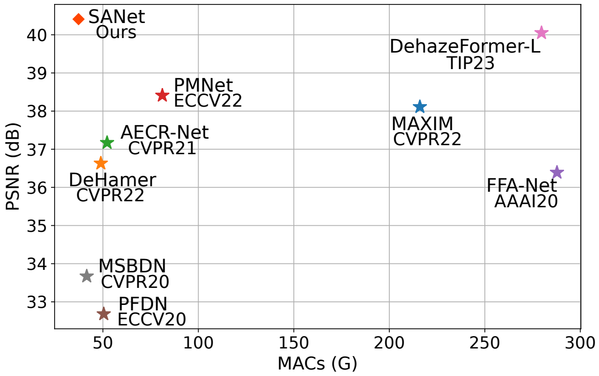


Figure 1: Accuracy and complexity comparisons between previous leading dehazing methods and ours SANet on the SOTS-Indoor [Li *et al.*, 2018] dataset. Our model receives a better performance while being computationally efficient.

More recently, Transformer model borrowed from natural language processing has shown state-of-the-art performance on high-level vision tasks. The core element, self-attention mechanism, is capable of modeling long-range dependencies effectively. However, its quadratic complexity with respect to the spatial resolution makes it infeasible for image restoration, which always involves high-resolution images, *e.g.*, 1680×1120 image size for defocus deblurring in DPDD [Abuolaim and Brown, 2020]. To alleviate this issue, many measures have been taken to improve efficiency in the realm of image restoration. For instance, a few methods restrict the operation region of self-attention to reduce complexity [Liang *et al.*, 2021; Wang *et al.*, 2022]. Restormer [Zamir *et al.*, 2022] applies self-attention among the channel dimension rather than the spatial dimension. Stripformer [Tsai *et al.*, 2022] develops strip-type self-attention for image deblurring. Though these methods realize the goal of reducing complexity to some extent, they do not break the nature of self-attention, *i.e.*, they still have quadratic complexity to the size of windows, channels, or strips.

In this paper, we exploit a strip attention mechanism for image restoration to harvest contextual information and meanwhile maintain high efficiency. Concretely, for each pixel, we perform information aggregation from its adjacent pixels in the same horizontal or vertical direction. This process is guided by the weights generated by convolutional layers. With joint horizontal and vertical aggregation, each location can implicitly receive information from a large region centered at itself. Furthermore, to enhance feature representation learning, we empirically adopt distinct receptive fields in feature groups to deal with degradation blurs of different sizes.

Our strip attention module has several key advantages. Firstly, by disintegrating attention into two directions, it significantly improves efficiency and can achieve large-scale receptive fields with negligible introduced computational complexity. Secondly, compared to the static filters of convolution operators, it is content-aware to adapt to the different input and blur. Thirdly, it is capable of capturing multi-scale contextual information. Our design is distinguished from other strip-type attention approaches. Specifically, CCNet [Huang *et al.*, 2019] utilizes recurrent criss-cross attention to capture full-image dependencies for semantic segmentation. CSWin transformer [Dong *et al.*, 2022] and Stripformer [Tsai *et al.*, 2022] execute self-attention within the strip-shaped regions in different directions. The attention weights of these methods are produced by matrix multiplication or affinity operation, which entails quadratic complexity. Differently, we generate weights from a simple bypass network and conduct integration in a cheap manner. Moreover, we exploit multi-scale receptive fields to boost performance.

Equipped with the proposed strip attention module, our SANet performs favorably against state-of-the-art algorithms on several image restoration tasks. For dehazing, as shown in Figure 1, SANet outperforms PMNet [Ye *et al.*, 2022] by 2.99 dB on the SOTS-Indoor [Li *et al.*, 2018] benchmark with 54% fewer MACs. For the defocus blur removal, SANet obtains 26.29 dB PSNR on DPDD [Abuolaim and Brown, 2020], an improvement of 0.31 dB over the strong Transformer model Restormer [Zamir *et al.*, 2022]. Our model also displays the potential on the desnowing task, surpassing NAFNet [Chen *et al.*, 2022] by 1.26 dB on CSD [Chen *et al.*, 2021].

The main contributions of the paper are as follows:

- We propose a strip attention module for image restoration that integrates multi-scale contextual information efficiently by performing horizontal and vertical local attention successively.
- Based on the proposed strip attention module, we establish SANet that performs favorably against state-of-the-art algorithms on several image restoration tasks.

2 Related Work

2.1 Image Restoration

Since image restoration plays an important role in photography, self-driving techniques, and medical imaging, it has drawn substantial attention from the industrial community and academia. This inverse problem has an ill-posed nature. To constrain the solution space, a flurry of conventional methods have been developed based on various assumptions and hand-crafted features [Zhang *et al.*, 2022]. Lately, the data-driven CNN-based frameworks have significantly advanced the performance of image restoration [Ren *et al.*, 2016; Ren *et al.*, 2018; Cui *et al.*, 2023a]. Among these networks, the U-shaped architecture [Ronneberger *et al.*, 2015] is a popular solution for hierarchical feature representation learning. Besides, numerous advanced modules have been created or borrowed from high-level tasks, including dilated convolution [Son *et al.*, 2021], skip connection [Liu *et al.*, 2019b], and multifarious attention mechanisms [Qin *et al.*, 2020]. More recently, Transformer models have been introduced into low-level tasks to help model long-range dependencies [Liang *et al.*, 2021].

2.2 Attention Mechanism

Attention mechanisms have been widely used in the computer vision community. In the context of image restoration, a great number of attention modules have been developed to capture inter-dependencies along channels [Liu *et al.*, 2019a; Zamir *et al.*, 2022], spatial coordinates [Zamir *et al.*, 2021], or both [Chen *et al.*, 2023]. For instance, FFA-Net [Qin *et al.*, 2020] leverages channel attention and pixel attention to deal with different types of information flexibly. GridDehazeNet [Liu *et al.*, 2019a] utilizes channel-wise attention to adjust the contributions of different streams for feature fusion. MPRNet [Zamir *et al.*, 2021] leverages the supervised attention module for feature filtering. These attention modules have boosted the performance of image restoration tasks.

Another line of this topic is to devise efficient self-attention for image restoration. Specifically, resembling Swin Transformer [Liu *et al.*, 2021], Uformer [Wang *et al.*, 2022] and SwinIR [Liang *et al.*, 2021] apply self-attention within local regions. Restormer [Zamir *et al.*, 2022] switches its focus from spatial dimensionality to channel self-attention. Stripformer [Tsai *et al.*, 2022] develops interlaced intra-strip and inter-strip attention layers for motion blur removal. However, these remedies still have quadratic complexity to the size of the region, channel, or strip.

In this paper, we present an ingenious strip attention module that performs efficient information integration in horizontal and vertical directions successively. Compared to the convolution operator, our paradigm not only inherits its high efficiency but also produces dynamic aggregation weights and an enlarged receptive field.

3 Methodology

In this section, we first describe the strip attention operation and then present the strip attention module. Next, we delineate the architecture of SANet for image restoration. Finally, we introduce the loss functions used for training.

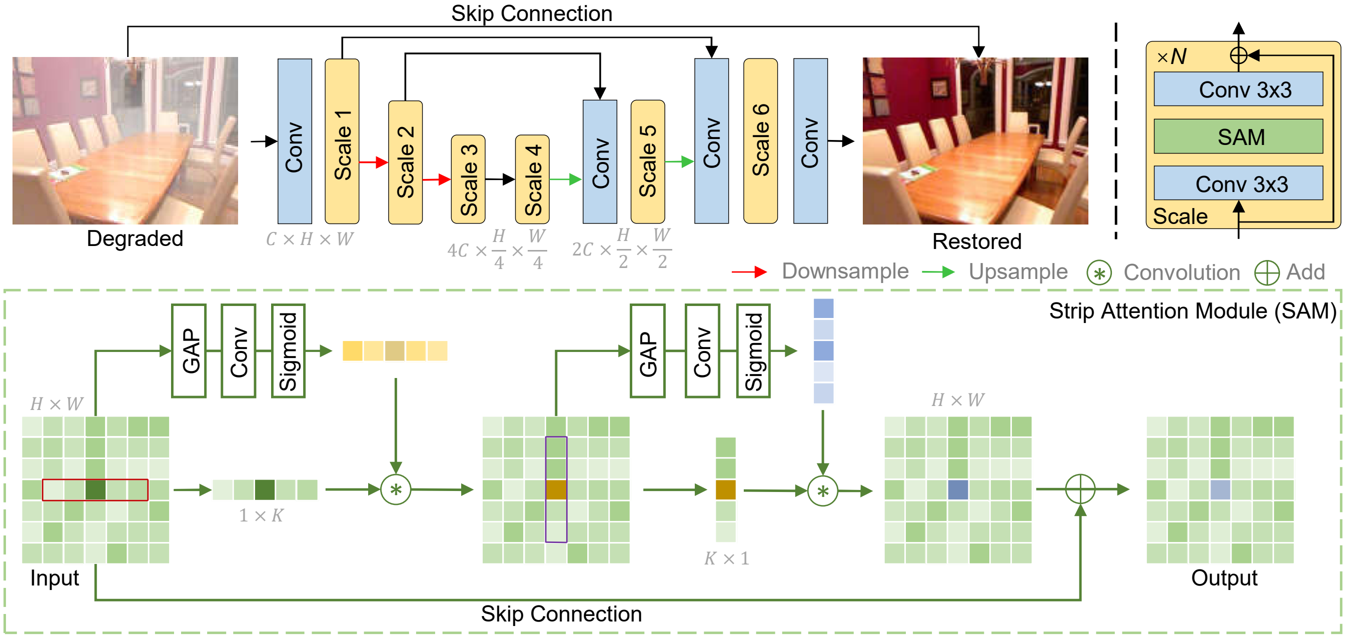


Figure 2: The architecture of SANet. **Top:** The overall pipeline. **Bottom:** The proposed strip attention module. We omit the channel dimension for clarity. The strip attention module only exists in the last residual block of each stage.

3.1 Strip Attention

Our main goal is to exploit a unit that can perform information integration efficiently and effectively. Before describing the formulation of the proposed strip attention, we first provide complexity analyses of self-attention.

Self-Attention

Self-attention has achieved successful stories in high-level vision tasks. However, due to its quadratic complexity, it is infeasible for image restoration tasks that always involve high-resolution images. Formally, give an input tensor $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where $H \times W$ denotes spatial coordinates and C is the number of channels, self-attention can be expressed as,

$$\begin{aligned} &\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Softmax}(\mathbf{Q}\mathbf{K}^{\top})\mathbf{V}, \\ &\text{where } \ \mathbf{Q} = \mathbf{X}\mathbf{W}^{Q}, \mathbf{K} = \mathbf{X}\mathbf{W}^{K}, \mathbf{V} = \mathbf{X}\mathbf{W}^{V} \end{aligned} \quad (1)$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{HW \times C}$, which are generated by using corresponding projection matrices ($\mathbf{W}^{Q}$, $\mathbf{W}^{K}$, and $\mathbf{W}^{V}$) and reshaping. We omit the normalization term for simplicity.

From Eq. 1, we can observe that the complexity of self-attention comes from three aspects: **(a)** the production of *query* ($\mathbf{Q}$), *key* ($\mathbf{K}$), and *value* ($\mathbf{V}$) with the complexity of $3HWC^2$; **(b)** generation of the attention map based on key-query dot-product with the complexity of $(HW)^2C$; **(c)** the weighted summation process with the complexity of $(HW)^2C$. We can see that in the last two terms, the complexity is quadratic to the spatial size.

Strip Attention

We aim to devise an efficient operator for information aggregation from the perspective of reducing the complexities of the above-mentioned three steps. Here, we take the horizontal strip attention as an example. Concretely, given any input feature $\mathbf{X}\in \mathbb{R}^{H \times W \times C}$, we remove the procedure of producing $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, and instead directly yield the attention weights via an extremely lightweight branch that consists of global average pooling (GAP) followed by 1×1 convolution layer and Sigmoid function. This process can be formally expressed as,

$$\mathbf{A} = \sigma(W_{1\times1}(\text{GAP}(\mathbf{X}))) \quad (2)$$

where $W_{1\times1}$ is a 1×1 convolution layer and σ denotes the Sigmoid function. $\mathbf{A}\in \mathbb{R}^{K}$, where K specifies the length of the strip for integration. Note that we share the resulting attention weights across both spatial and channel dimensions for further efficiency.

Regarding the weighted sum operation, instead of operating on the whole image like self-attention or on the strip of size $n \times W$ ($n < H$) like Stripformer [Tsai *et al.*, 2022] and CSWin Transformer [Dong *et al.*, 2022], we execute our information integration within the strip of size $1 \times K$ ($K < W$) based on the obtained attention weights, which can be formally expressed as,

$$\hat{\mathbf{X}}_{h,w,c} = \sum_{k=0}^{K-1} \mathbf{A}_k \mathbf{X}_{h,w-\lfloor \frac{K}{2} \rfloor + k,c} \quad (3)$$

Rather than generating the attention weights with a similar shape to that of self-attention and then performing integration via matrix multiplication, inspired by [Zhou *et al.*, 2021], we adopt a more reasonable convolution-type integration as shown in Figure 3 (c), where each pixel receives information from the region centered at itself.

To summarize, our strip attention operator can be formally expressed as:

$$\hat{\mathbf{X}} = \mathcal{S}_K(\mathbf{X}) \quad (4)$$

3.2 Strip Attention Module

It has been illustrated in prior works [Zamir *et al.*, 2022; Wang *et al.*, 2022] that enlarging the receptive field of the network is beneficial to image restoration. Motivated by this fact, we present an efficient manner to expand the receptive field of each pixel by exploiting the above-mentioned strip attention operator. Specifically, we develop a strip attention module that carries out strip attention operations in both vertical and horizontal directions to harvest long-range contexts, as shown in the bottom part of Figure 2. Furthermore, we combine different K within each attention operation to pursue multi-scale receptive fields. More concretely, with the input tensor $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, we first divide it into two parts by splitting the channel dimension evenly and then impose the horizontal strip attention on each part separately with different strip lengths K. Next, we perform the multi-scale strip attention in the vertical direction. The final output is produced by adding the original input $\mathbf{X}$. The entire process of the proposed strip attention module can be formally expressed as:

$$\mathbf{Y} = [\mathcal{S}^{Ve}_{k_1}(\mathcal{S}^{Ho}_{k_1}(\mathbf{X}_1)), \mathcal{S}^{Ve}_{k_2}(\mathcal{S}^{Ho}_{k_2}(\mathbf{X}_2))] + \mathbf{X} \tag{5}$$

where $\mathcal{S}^{Ho}$ and $\mathcal{S}^{Ve}$ denote the horizontal attention and vertical attention, respectively; $[\cdot,\cdot]$ is concatenation; $\mathbf{X}_1$ and $\mathbf{X}_2$ are obtained by splitting the feature on channel dimension evenly. Our strip attention module implicitly enlarges the receptive field of the network. As shown in Figure 3, the horizontal and vertical strip attention perform information integration in two directions, respectively. For convenience, we only pick a few representative pixels for illustration. The horizontal one gives $\mathrm{B} = w_{AB}\mathrm{A} + w_{BB}\mathrm{B} + w_{CB}\mathrm{C}$, where w denotes the attention weight. By using two-directional strip attention successively, the value of pixel D in Figure 3 (b) is computed by:

$$\begin{aligned} \mathrm{D} &= w_{BD}\mathrm{B} + w_{DD}\mathrm{D} \\ &= w_{BD}(w_{AB}\mathrm{A} + w_{BB}\mathrm{B} + w_{CB}\mathrm{C}) + w_{DD}\mathrm{D}. \end{aligned} \tag{6}$$

As a consequence, the pixel in the center receives contexts from the whole region determined by K.

3.3 Overall Architecture

The overall pipeline of the proposed SANet is illustrated in Figure 2 (Top). SANet adopts the popular encoder-decoder architecture to learn hierarchical representations efficiently and consists of six scales in total. Specifically, given a degraded image with the shape of $\mathbb{R}^{3\times H\times W}$, a single convolution layer is utilized to generate the shallow feature map of size $\mathbb{R}^{C\times H\times W}$. Then, the resulting feature is fed into the encoder layers (Scale 1-3). In this process, the number of channels is expanded, while the spatial resolution is reduced gradually from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{4C\times \frac{H}{4}\times\frac{W}{4}}$. Each stage contains a stack of residual blocks, and the last one involves the proposed strip attention module. The downsampling operation is accomplished by the strided convolution. Next, the feature with the lowest resolution passes through the decoder layers (Scale 4-6) to recover the high-resolution representations progressively. For feature upsampling, we adopt the transposed convolution. To alleviate the issue of information loss caused by downsampling, we apply the feature-level skip connections as previous works [Zamir *et al.*, 2022; Wang *et al.*, 2022]. Concretely, the encoder features are concatenated with the corresponding decoder features, followed by a convolution layer to adjust the channel dimension. The final sharp image is produced by adding the original input image, which forces the network to focus only on the residual information learning. Besides, to ease the training difficulty, multi-input and multi-output strategies are adopted following recent methods [Cho *et al.*, 2021; Mao *et al.*, 2021].

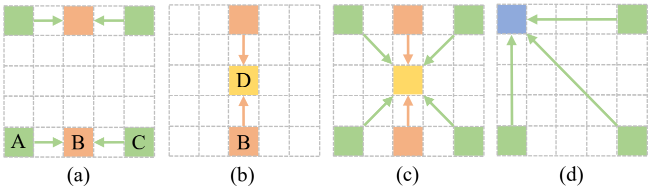


Figure 3: Signal integration paradigm of our strip attention and self-attention. **(a)** Horizontal strip attention operator. **(b)** Vertical strip attention operator. **(c)** Strip attention module. **(d)** Self-attention.

3.4 Loss Functions

To facilitate feature refinement in spatial and frequency domains simultaneously, we use the dual-domain L$_1$ loss [Cho *et al.*, 2021] to train our network. For each output, the loss function is given by:

$$\begin{aligned} L_s &= \frac{1}{S}\|\hat{\mathbf{I}} - \mathbf{I}\|, \\ L_f &= \frac{1}{S}\|\mathcal{F}(\hat{\mathbf{I}}) - \mathcal{F}(\mathbf{I})\|, \\ L &= L_s + \lambda L_f \end{aligned} \tag{7}$$

where $\hat{\mathbf{I}}, \mathbf{I}$ are the predicted image and ground-truth, respectively; S depicts the total elements for normalization; and $\mathcal{F}$ is the fast Fourier transform (FFT). λ is set to 0.1.

4 Experiments

To verify the effectiveness of our SANet, we conduct extensive experiments on several image restoration tasks, including single-image defocus deblurring (DPDD [Abuolaim and Brown, 2020]), image dehazing (RESIDE [Li *et al.*, 2018]), and image desnowing (CSD [Chen *et al.*, 2021]). In the following, we first introduce the training settings, and then we report our results on the above datasets. Finally, we carry out a series of ablation experiments.

4.1 Implementation Details

We train the proposed network via Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is set to $1e^{-4}$ and reduced to $1e^{-6}$ gradually with the cosine annealing. The batch size is set as 8 for the RESIDE-Outdoor [Li *et al.*, 2018] dataset and 4 for others. Models are trained on the patch size of 256×256. We adopt only horizontal flips for data augmentation. We choose $k_1 = 7$ and $k_2 = 11$ in Eq. 5. According to the task complexity, we deploy varying numbers of residual blocks N in each scale for different tasks, *i.e.*, $N = 4$ for image dehazing and desnowing, and $N = 16$ for image defocus deblurring.

	SOTS-Indoor		SOTS-Outdoor		Overhead	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑	Params (M)	MACs (G)
DCP [He *et al.*, 2010]	16.62	0.818	19.13	0.815	-	-
GCANet [Chen *et al.*, 2019]	30.23	0.980	-		0.702	18.41
GridDehazeNet [Liu *et al.*, 2019a]	32.16	0.984	30.86	0.982	0.956	21.49
MSBDN [Dong *et al.*, 2020]	33.67	0.985	33.48	0.982	31.35	41.54
PFDN [Dong and Pan, 2020]	32.68	0.976	-		11.27	50.46
FFA-Net [Qin *et al.*, 2020]	36.39	0.989	33.57	0.984	4.456	287.8
AECR-Net [Wu *et al.*, 2021]	37.17	0.990	-		2.611	52.20
MAXIM [Tu *et al.*, 2022]	38.11	0.991	34.19	0.985	14.1	108
DeHamer [Guo *et al.*, 2022]	36.63	0.988	35.18	0.986	132.45	48.93
PMNet [Ye *et al.*, 2022]	38.41	0.990	34.74	0.985	18.90	81.13
DehazeFormer-L [Song *et al.*, 2022]	40.05	**0.996**	-		25.44	279.7
SANet (Ours)	**40.40**	**0.996**	**38.01**	**0.995**	3.81	37.26

Table 1: Image dehazing results on SOTS [Li *et al.*, 2018]. SANet receives higher scores with fewer MACs than most competitors.

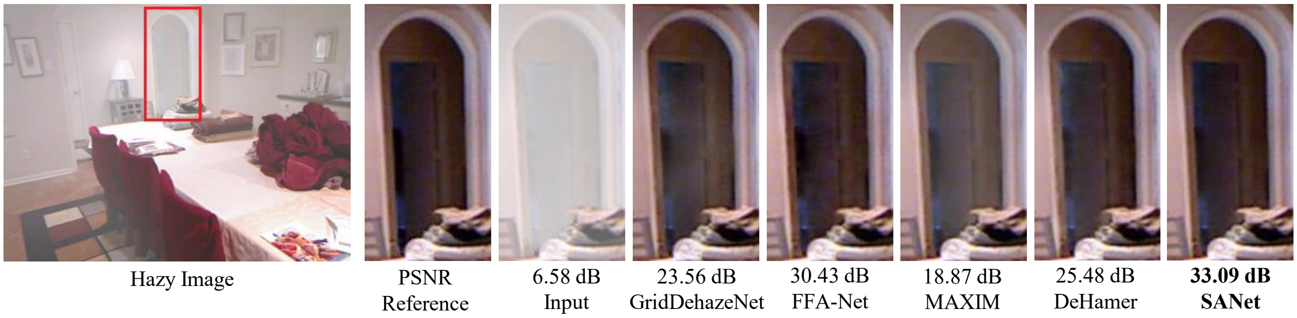


Figure 4: Image dehazing comparisons on the SOTS-Indoor [Li *et al.*, 2018] dataset among GridDehazeNet [Liu *et al.*, 2019a], FFA-Net [Qin *et al.*, 2020], MAXIM [Tu *et al.*, 2022], DeHamer [Guo *et al.*, 2022], and our SANet. Our model is more effective in haze removal.

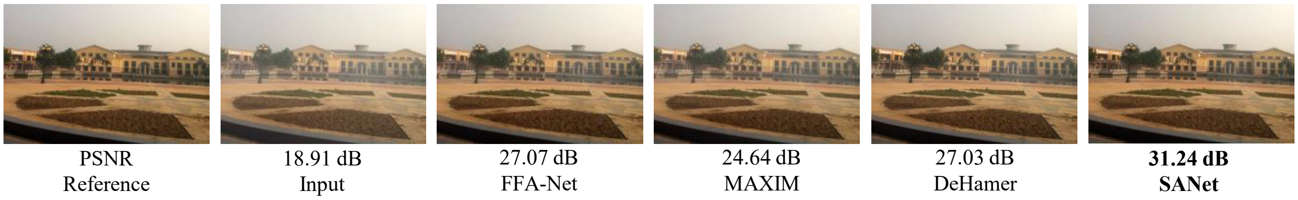


Figure 5: Image dehazing comparisons on the SOTS-Outdoor [Li *et al.*, 2018] dataset among FFA-Net [Qin *et al.*, 2020], MAXIM [Tu *et al.*, 2022], DeHamer [Guo *et al.*, 2022], and our SANet.

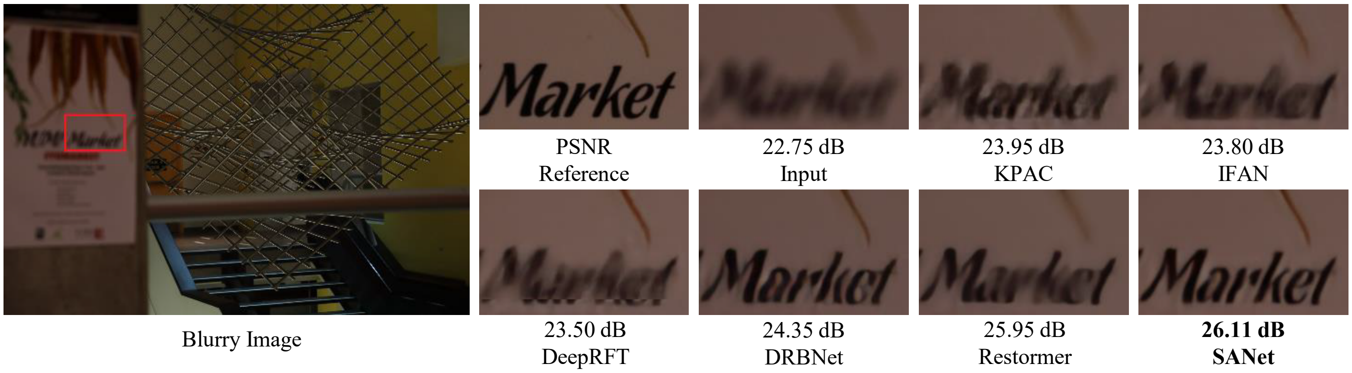
4.2 Main Results

Image dehazing. We train the network on the RESIDE [Li *et al.*, 2018] dataset and test on the SOTS [Li *et al.*, 2018] dataset. The results are reported in Table 1. Our SANet achieves better performance with lower complexity than most approaches. Particularly on the SOTS-Outdoor dataset, SANet yields a 2.83 dB performance gain over the expensive Transformer model DeHamer [Guo *et al.*, 2022] with only 76% MACs and 3% parameters. Compared to the recent algorithm DehazeFormer-L [Song *et al.*, 2022], our model surpasses it by 0.35 dB in terms of PSNR on SOTS-Indoor, while having 6.68× fewer parameters and 7.5× fewer MACs. The qualitative comparisons on the SOTS-Indoor and SOTS-Outdoor datasets are exhibited in Figure 4 and Figure 5, respectively. We can see that SANet is more effective in removing haze blur, and the images produced by our model are visually closer to the target ones than other algorithms.

Single-image defocus deblurring. We compare image fidelity scores of our method with both learning-based single-image defocus deblurring methods and conventional ones, *e.g.*, JNB [Shi *et al.*, 2015] and EBDB [Karaali and Jung, 2018], on the DPDD [Abuolaim and Brown, 2020] dataset. The comparison results in Table 2 show that our model outperforms the strong Transformer model Restormer [Zamir *et al.*, 2022] in most cases. Particularly in the indoor scene category, SANet produces a substantial gain of 0.43 dB over Restormer. Furthermore, our method outperforms DRBNet [Ruan *et al.*, 2022] by 0.56 dB PSNR on the combined

	Indoor Scenes				Outdoor Scenes				Combined			
Method	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓
EBDB [Karaali and Jung, 2018]	25.77	0.772	0.040	0.297	21.25	0.599	0.058	0.373	23.45	0.683	0.049	0.336
DMENet [Lee *et al.*, 2019]	25.50	0.788	0.038	0.298	21.43	0.644	0.063	0.397	23.41	0.714	0.051	0.349
JNB[Shi *et al.*, 2015]	26.73	0.828	0.031	0.273	21.10	0.608	0.064	0.355	23.84	0.715	0.048	0.315
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**
SANet (Ours)	**29.30**	0.878	**0.024**	0.163	**23.43**	**0.748**	**0.049**	0.227	**26.29**	**0.811**	**0.037**	0.196

Table 2: Single-image defocus deblurring results on the DPDD [Abuolaim and Brown, 2020] dataset.



Blurry Image

PSNR
Reference

22.75 dB
Input

23.95 dB
KPAC

23.80 dB
IFAN

23.50 dB
DeepRFT

24.35 dB
DRBNet

25.95 dB
Restormer

26.11 dB
SANet

Figure 6: Single-image defocus deblurring comparisons on the DPDD [Abuolaim and Brown, 2020] dataset among KPAC [Son *et al.*, 2021], IFAN [Lee *et al.*, 2021], DeepRFT [Mao *et al.*, 2021], DRBNet [Ruan *et al.*, 2022], Restormer [Zamir *et al.*, 2022], and our SANet. Our model recovers more faithful details than other methods.

category. The visual results in Figure 6 illustrate that the proposed network recovers more faithful details than other competitive frameworks.

Image desnowing. The desnowing comparisons on the CSD [Chen *et al.*, 2021] dataset are provided in Table 3. We can see that our method obtains higher scores than other approaches. Compared to the recent algorithm NAFNet [Chen *et al.*, 2022], SANet provides a performance boost of 1.26 dB PSNR. Furthermore, our model shows a 2.64 dB improvement over the Transformer model MSP-Former [Chen *et al.*, 2023]. Visual results presented in Figure 7 show that our SANet generates a cleaner image than other algorithms.

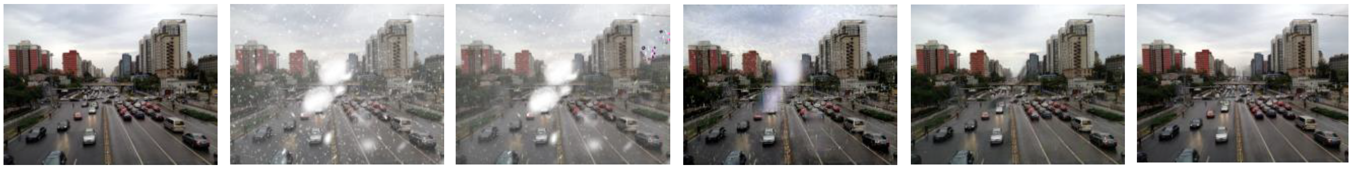
Method	PSNR	SSIM
DesnowNet [Liu *et al.*, 2018]	20.13	0.81
CycleGAN [Engin *et al.*, 2018]	20.98	0.80
All in One [Li *et al.*, 2020]	26.31	0.87
JSTASR [Chen *et al.*, 2020]	27.96	0.88
HDCW-Net [Chen *et al.*, 2021]	29.06	0.91
TransWeather [Valanarasu *et al.*, 2022]	31.76	0.93
MSP-Former [Chen *et al.*, 2023]	33.75	0.96
NAFNet [Chen *et al.*, 2022]	35.13	0.97
SANet (Ours)	**36.39**	**0.98**

Table 3: Image desnowing results on CSD [Chen *et al.*, 2021]. SANet outperforms other methods significantly.

4.3 Ablation Studies

For ablation experiments, we study diverse design choices for the strip attention module, including the combination pattern of strip attention, different strip lengths, and activation functions. Furthermore, we compare our module with other attention units and depth-wise convolution to demonstrate the effectiveness of our method. To this end, we train SANet on the dehazing task with the RESIDE-Indoor [Li *et al.*, 2018] dataset. Unless specified otherwise, the hyperbolic tangent function serves as the activation function in Eq. 2, and we only adopt the single receptive field for the strip attention module with $K = 5$. The training configurations are consistent with the main experiment except that N is set to 1. MACs are computed on the size of 256 × 256. The baseline model is obtained by removing the proposed attention module from our model.

Improvements of strip attention module. Table 4 shows that two-directional strip attention units both produce favorable gains over the baseline model with negligible introduced parameters and complexity. Using two strip attention operators in different directions leads to further accuracy improvement. Especially for the horizontal-vertical version, our model achieves a gain of 3.39 dB over the baseline, while only consuming additional 0.04 M parameters and 0.07 G



PSNR Reference	13.30 dB Input	14.30 dB DesnowNet	21.01 dB JSTASR	27.59 dB HDCW-Net	**32.28 dB** **SANet**

Figure 7: Image desnowing comparisons on the CSD [Chen *et al.*, 2021] dataset among DesnowNet [Liu *et al.*, 2018], JSTASR [Chen *et al.*, 2020], HDCW-Net [Chen *et al.*, 2021], and our SANet.

Method	PSNR	Params (M)	MACs (G)
Baseline	31.33	1.48	15.44
Horizontal strip	34.19	1.50	15.48
Vertical strip	33.93	1.50	15.48
Parallel	34.53	1.52	15.51
Vertical-Horizontal	34.36	1.52	15.51
Horizontal-Vertical	34.72	1.52	15.51

Table 4: Ablation studies for strip attention module. *Parallel* variant combines the outcomes of two-directional strip attention operators via addition.

Method	Softmax	Tanh	Sigmoid
PSNR	34.35	34.36	34.68

Table 5: Different activation functions.

Method	Self-attention	Window attention	Ours
PSNR/MACs (G)	34.48/20	34.44/16.71	35.85/15.58

Table 6: Comparisons with other attention modules.

MACs, illustrating the effectiveness of our design.

Different receptive fields. We further exploit the impact of the receptive field by changing the strip size in the strip attention module. The results are shown in Figure 8. As the increase of the strip length, we can observe a consistent improvement in terms of PSNR. Our model receives a remarkable gain of 1.36 dB when the receptive field is enlarged from 3 to 11, while only introducing 0.12 G MACs. Furthermore, to deal with blurs of various sizes, we adopt the multi-scale receptive fields, *i.e.*, 7 and 11, as we elaborate in Sec. 3.2. This strategy leads to 35.45 dB PSNR, 0.05 dB and 0.14 dB higher than a single kernel 11 and 7, receptively.

Design choices for activation function. Instead of inheriting the Softmax function from the canonical self-attention, we explore more choices in Table 5 based on the vertical-horizontal variant in Table 4. The Sigmoid version obtains higher accuracy than Softmax by breaking the sum-to-one property and is 0.32 dB higher than that of the Tanh version.

Comparisons with alternatives. As our strip attention module implicitly receives the same receptive field as the depth-wise convolution when K is equal to the kernel size of the latter, we compare our module with depth-wise convolution in Figure 8. With the same receptive field, our model consistently outperforms the depth-wise convolution version with fewer extra complexities. To further verify the superiority of our method, we provide comparisons between other self-attention units and our module. We can see from Table 6 that, with the best choices of receptive field, activation, and combination order of strip attention units, our final design is superior to the global self-attention and the window-based variant in terms of accuracy and computation overhead. Due to the large complexities of global self-attention, we only insert it into scale 3-4, which have the lowest resolution.

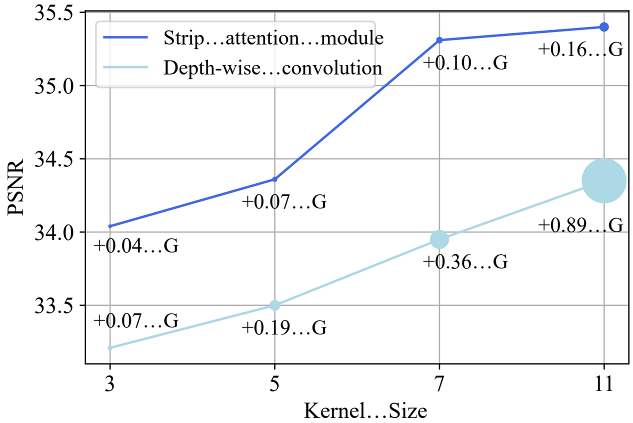


Figure 8: Ablations on receptive fields. For the strip attention module and depth-wise convolution, we can observe a consistent PSNR improvement when increasing the receptive field size. Our method is more efficient than the depth-wise convolution. The annotated number and dot size indicate the introduced MACs over the baseline.

5 Conclusion

In this paper, we develop a novel image restoration model that is computationally efficient in integrating contexts for feature representation enhancement. Specifically, our strip attention unit realizes efficient information aggregation by modifying the three steps of self-attention, while maintaining the content-aware property based on the learned attention weights. Furthermore, the proposed strip attention module enlarges the receptive field by combining two-directional strip attention units and adopts multi-scale kernels to well handle blurs with various sizes. Comprehensive experiments on several image restoration tasks demonstrate that SANet performs favorably against state-of-the-art algorithms.

References

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

[Chen *et al.*, 2019] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *2019 IEEE Winter Conference on Applications of Computer Vision*, pages 1375–1383, 2019.

[Chen *et al.*, 2020] 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. Springer, 2020.

[Chen *et al.*, 2021] 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.

[Chen *et al.*, 2022] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, 2022.

[Chen *et al.*, 2023] Sixiang Chen, Tian Ye, Yun Liu, Taodong Liao, Jingxia Jiang, Erkang Chen, and Peng Chen. Msp-former: Multi-scale projection transformer for single image desnowing. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023.

[Cho *et al.*, 2021] 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, October 2021.

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

[Cui *et al.*, 2023b] Yuning Cui, Yi Tao, Wenqi Ren, and Alois Knoll. Dual-domain attention for image deblurring. In *Association for the Advancement of Artificial Intelligence*, 2023.

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

[Dong *et al.*, 2020] 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*, June 2020.

[Dong *et al.*, 2022] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022.

[Engin *et al.*, 2018] 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*, June 2018.

[Guo *et al.*, 2022] 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.

[He *et al.*, 2010] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:2341–2353, 2010.

[Huang *et al.*, 2019] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 603–612, 2019.

[Karaali and Jung, 2018] Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27:1126–1137, 2018.

[Lee *et al.*, 2019] Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[Lee *et al.*, 2021] 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.

[Li *et al.*, 2018] 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.

[Li *et al.*, 2020] 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*, June 2020.

[Liang *et al.*, 2021] 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*, pages 1833–1844, 2021.

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

[Liu *et al.*, 2019a] 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.

[Liu *et al.*, 2019b] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2019.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[Mao *et al.*, 2021] 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.

[Qin *et al.*, 2020] 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*, volume 34, pages 11908–11915, 2020.

[Ren *et al.*, 2016] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang. Single image dehazing via multi-scale convolutional neural networks. In *European Conference on Computer Vision*, pages 154–169. Springer, 2016.

[Ren *et al.*, 2018] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[Ruan *et al.*, 2022] 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.

[Shi *et al.*, 2015] Jianping Shi, Li Xu, and Jiaya Jia. Just noticeable defocus blur detection and estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2015.

[Son *et al.*, 2021] 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.

[Song *et al.*, 2022] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *arXiv preprint arXiv:2204.03883*, 2022.

[Tsai *et al.*, 2022] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European Conference on Computer Vision*, 2022.

[Tu *et al.*, 2022] 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.

[Valanarasu *et al.*, 2022] 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, June 2022.

[Wang *et al.*, 2022] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022.

[Wu *et al.*, 2021] 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, June 2021.

[Ye *et al.*, 2022] 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. Springer, 2022.

[Zamir *et al.*, 2021] 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.

[Zamir *et al.*, 2022] 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.

[Zhang *et al.*, 2022] 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.

[Zhou *et al.*, 2021] Jingkai Zhou, Pichao Wang, Fan Wang, Qiong Liu, Hao Li, and Rong Jin. Elsa: Enhanced local self-attention for vision transformer. *arXiv preprint arXiv:2112.12786*, 2021.