PSNet: Towards Efficient Image Restoration With Self-Attention

Yuning Cui and Alois Knoll

***Abstract*—Image restoration aims to recover a clean image from various degradations, *e.g.,* haze, snow, and blur, playing an important role in robot vision, autonomous vehicles, and medical imaging. Recently, the use of Transformer has witnessed a significant improvement in multifarious image restoration tasks. However, despite a few remedies to reduce the quadratic complexity of self-attention, these approaches are still impractical for real-world applications, which need high efficiency and speed. To ameliorate this issue, we propose an efficient framework for image restoration based on self-attention. To this end, we combine the strengths of patch-based and strip-based self-attention units to improve efficiency. More specifically, we apply self-attention of different operation scales to features of different resolutions, i.e., we adopt a relatively smaller region for self-attention on high-resolution features while a larger region for low-restoration features. In addition, instead of using global self-attention in each partitioned region, we leverage a strip-based version for low complexity. To further improve efficiency, we insert our design into a U-shaped CNN network to establish our framework, dubbed PSNet. Extensive experiments demonstrate that our network receives state-of-the-art performance on five representative image restoration tasks with low computational complexity and high speed, including single-image defocus deblurring, image dehazing, image motion deblurring, image desnowing, and image denoising.**

***Index Terms*—Deep learning for visual perception, visual learning, representation learning.**

I. Introduction

DUE to the physical limitation of devices or bad weather, the captured images often suffer from various degradations, *e.g.,* snowflake, haze, and noise. Such degradations often degrade the visibility of images and affect the performance of high-level tasks, such as object detection and segmentation. In this regard, image restoration is immensely conducive by removing those undesired degradations and reconstructing the latent clean image. Towards this ill-posed problem, many conventional algorithms have been proposed by using various hand-crafted features. However, these methods are not robust enough to apply to the more complicated real-world scenarios [1], [2].

With the rapid development of convolutional neural networks (CNNs), a multitude of data-driven CNN-based networks have been developed, which have a strong ability to learn generalizable priors from large-scale collected datasets. Despite the improved performance, the convolution operator has two main defects. Firstly, it has a limited receptive field, which is disadvantageous to manage large-scale degradation blurs. Secondly, the fixed kernel values are not flexible enough to remove non-uniform blurs and adapt to inputs [2].

To ameliorate the above-mentioned issues, Transformer models have been introduced into low-level tasks and have significantly advanced the performance of image restoration tasks, *e.g.,* image dehazing [5], deblurring [6], and desnowing [7]. Despite a few remedies to reduce the inherited quadratic computational complexity, these frameworks still have heavy computation overhead and are not applicable to practical applications. For example, Restormer [6] takes 1.218 seconds to process a 720×1280 blurry image.[1] DehazeFormer-L [1] obtains 40.05 dB PSNR on the dehazing dataset SOTS-Indoor [3] while having 279.7 G FLOPs, which is about five times more than the CNN-based AECR-Net[2] [8].

In this study, we propose an efficient image restoration framework, dubbed PSNet, which is established on patch-based and strip-based self-attention units. We elaborately design the region size for self-attention to achieve our goal. Concretely, we apply the relatively large attention operation region to low-resolution features while we use a small region for high-resolution features. To further improve efficiency, instead of performing global self-attention in each region as window-based self-attention [9], the two-direction strip-based self-attention is leveraged for information aggregation.

Incorporating our design into a U-shaped backbone, our simple convolutional framework achieves state-of-the-art performance on several image restoration tasks. The main contributions of this study can be summarized as follows:

- We propose an efficient image restoration framework, named PSNet, which enhances representation learning while remaining efficient by elaborately determining the operation region size for self-attention.
- The proposed network achieves state-of-the-art performance on 13 benchmark datasets among five image restoration tasks, including single-image defocus deblurring, image dehazing, image desnowing, image motion deblurring, and image denoising.

Manuscript received 1 March 2023; accepted 7 July 2023. Date of publication 31 July 2023; date of current version 7 August 2023. This letter was recommended for publication by Associate Editor M. Burke and Editor A. Faust upon evaluation of the reviewers' comments. *(Corresponding author: Yuning Cui.)*

The authors are with the School of Computation, Information and Technology, Technical University of Munich, 85748 Munich, Germany (e-mail: yuning.cui@in.tum.de; knoll@in.tum.de).

Digital Object Identifier 10.1109/LRA.2023.3300254

[1]Inference time is computed on an NVIDIA Tesla V100 GPU with Intel Xeon Platinum 8255 C CPU.

[2]FLOPs are measured on 256×256 patches.

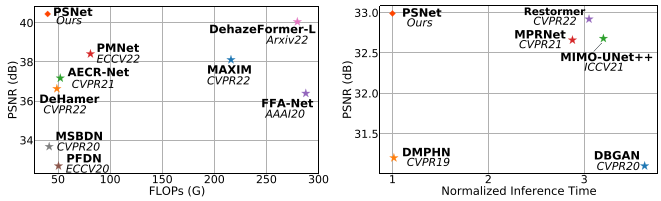


Fig. 1. Comparisons between PSNet and other state-of-the-art algorithms. **Left**: PSNR *vs.* FLOPs on the SOTS-Indoor [3] dataset for image dehazing. **Right**: PSNR *vs.* normalized inference time (by ours) on the GoPro [4] dataset for image motion deblurring.

II. Related Work

Image restoration aims to recover a clean image from a degraded observation. Conventional methods are mostly based on various assumptions and hand-crafted features. Recent CNN-based methods have produced a promising performance for image restoration tasks with advanced designs, such as encoder-decoder architecture [10] and various attention units [11], [12]. For instance, Qin et al. [12] proposed an attention module by combining the channel attention and pixel attention mechanism for image dehazing. Cho et al. [10] developed a coarse-to-fine deblurring network based on multi-input and multi-output techniques. Lee et al. [13] presented the iterative adaptive convolution to handle spatially varying defocus blur. However, the convolution operator is incapable of providing large receptive fields and adapting to input features.

Recently, inspired by their success in high-level vision tasks, Transformer models have been introduced into image restoration to model long-range dependencies. To reduce the complexity of self-attention, a few methods have been proposed. Wang et al. [9] adopted the window-based self-attention by restricting the region of self-attention. Zamir et al. [6] switched self-attention from the spatial dimension to the channel dimension. However, this operation sacrifices the spatial modeling ability for high efficiency. Tsai et al. [14] proposed an efficient Transformer model using strip-based self-attention. In this article, we further reduce the complexity of strip-type self-attention by introducing different partition schemes according to the resolution of features.

III. Methodology

We first delineate the overall pipeline of PSNet. Next, we introduce the details of our efficient self-attention module. Finally, the used loss functions are introduced.

A. Overall Pipeline

The schematic of PSNet is illustrated in Fig. 2. As shown in Fig. 2(a), our network adopts the popular encoder-decoder solution to learn hierarchical features. The encoder and decoder sub-networks both contain three scales. Here, the scale refers to the resolution of features, and the first scale has the highest resolution. Specifically, given a degraded image of size $H \times W \times 3$, where $H \times W$ denotes spatial locations, and 3 represents the number of channels. PSNet first employs a 3×3 convolution to extract shallow features of size $H \times W \times C$. Then the shallow features are fed into the ResBlock and IA-S (Small) to obtain the output of the first scale. ResBlock is comprised of n residual blocks, which contain two 3×3 convolutional layers with a GELU activation function in between, as illustrated in Fig. 2(b). Next, the resulting features are down-sampled to the size of $\frac{H}{2} \times \frac{W}{2} \times 2C$ by a strided convolution. The deepest features are yielded after three scales of the encoder network.

To restore the clean image, the deepest features pass through the decoder network, which has a similar architecture to the encoder network. During this process, decoder features are concatenated with the encoder features, followed by 1×1 convolution to reduce channels by half. The final predicted sharp image is generated after the image-level skip connection. Up-sampling operation is implemented by transposed convolution.

Our efficient attention module (EAM) is inserted into a residual block to establish the information aggregation (IA) module, as represented in Fig. 2(c). Moreover, to inject more information about the input image into the network and provide more supervision signals, we apply the multi-input and multi-output strategies, following previous algorithms [10], [15], [16]. Specifically, the input of reduced resolution is merged into the mainstream via a simple convolution block and concatenation, followed by a convolution to adjust the number of channels. In addition, the decoder network outputs the low-resolution images after IA-L (Large) and IA-M (Medium) to assist in training.

B. Efficient Attention Module (EAM)

EAM, which involves our partition schemes and the self-attention operation, is used to perform information aggregation and improve representation learning for image restoration. The two-direction strip-based attention [14] carries out the self-attention operation while the partition part determines the region size for the former.

1) Strip-Based Self-Attention: Since the vertical and horizontal strip-based self-attention units share a similar paradigm, we take the horizontal self-attention (HSA) as an example, as illustrated in Fig. 2(d). Given any input $X^h \in \mathbb{R}^{H' \times W' \times \frac{C}{2}}$, we first split X^h into non-overlapping horizontal strips as $X_i^h, i \in \{1, 2, \ldots, H'\}$. Following previous algorithms [6], [14], we produce query (Q_{ij}^h), key (K_{ij}^h), and value (V_{ij}^h) with the multi-head mechanism as

$$(Q_{ij}^h, K_{ij}^h, V_{ij}^h) = (X_i^h W_j^Q, X_i^h W_j^K, X_i^h W_j^V) \tag{1}$$

where W_j^Q, W_j^K, and $W_j^V \in \mathbb{R}^{\frac{C}{2} \times \frac{C}{2d}}$, $j \in \{1, 2, \ldots, d\}$ denote projection matrices, and d is the number of heads. Next, the enhanced features can be obtained via multi-head attention, expressed as:

$$\hat{X}_{ij}^h = \text{Softmax}\left(\frac{Q_{ij}^h (K_{ij}^h)^\top}{\sqrt{C/2\,d}}\right) V_{ij}^h \tag{2}$$

Thus, the output features of HSA can be obtained as $\hat{X}^h \in \mathbb{R}^{H' \times W' \times \frac{C}{2}}$ by concatenating all heads over channel dimension and folding all horizontal strips among vertical direction. The whole process of HSA can be concluded as follows:

$$\hat{X^h} = \text{HSA}(X^h) \tag{3}$$

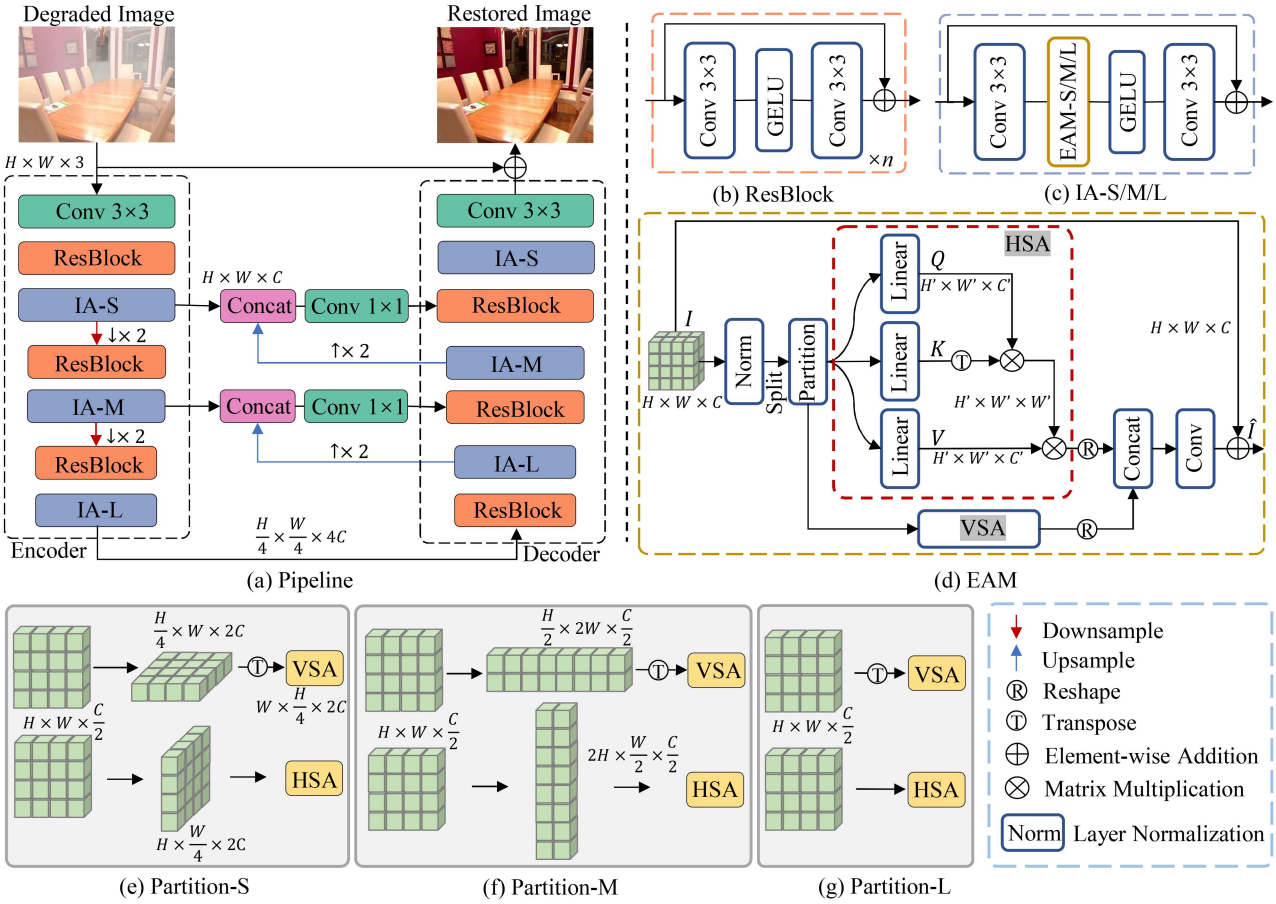


Fig. 2. Architecture of PSNet for image restoration. (a) The overall pipeline of PSNet. (b) ResBlock contains n residual blocks. (c) Information aggregation (IA) module with different configurations, i.e., EAM-S, EAM-M, and EAM-L. (d) Efficient self-attention module (EAM) mainly contains horizontal self-attention (HSA), vertical self-attention (VSA), and different Partition variants, i.e., Partition-S (e), Partition-M (f), and Partition-L (g).

Similarly, the output features of the vertical branch (VSA) can be yielded by:

$$\hat{X^v} = \mathrm{VSA}(X^v) \in \mathbb{R}^{W' \times H' \times \frac{C}{2}} \tag{4}$$

2) Partition: To improve efficiency, we partition features before passing them through strip attention. On the basis of the resolution of features, we introduce three Partition variants.

Partition-S (Small): As illustrated in Fig. 2(e), the input features have been split into two components of size $H \times W \times \frac{C}{2}$. To reduce the operation region of strip-based attention, the bottom part is first partitioned into four patches, which are then concatenated into the size of $\frac{H}{2} \times \frac{W}{2} \times 2\,C$. This process is not shown in the figure for simplicity. Next, the width of resulting patches is further reduced by half, and the size of reshaped features is $H \times \frac{W}{4} \times 2\,C$. During this process, each pixel can receive information from the corresponding location of other patches, which implicitly enlarges the receptive field. Similarly, the above process can be applied to the other half of features to obtain features of size $\frac{H}{4} \times W \times 2\,C$, which is transposed to $W \times \frac{H}{4} \times 2\,C$ before being fed into VSA.

Partition-M (Medium): Since the second scales of both encoder and decoder networks have relatively smaller features than that of the first scale, we enlarge the attention region to obtain large-scale receptive fields. As illustrated in Fig. 2(f), the height of the top features is reduced by half. The sizes of input features for VSA and HSA are $2W \times \frac{H}{2} \times \frac{C}{2}$ and $2H \times \frac{W}{2} \times \frac{C}{2}$, respectively. As illustrated in Fig. 3, the receptive fields have been enlarged to the half size of features in two directions.

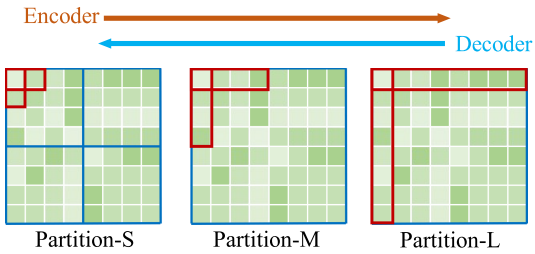


Fig. 3. Schematic of three Partition methods. The encoder network performs information aggregation in the order of a small region to a large region. The decoder network applies the opposite order to recover clean features in a coarse-to-fine manner.

Partition-L (Large): Partition-L is utilized in the third scale of both encoder and decoder networks, which have the smallest resolution features. For this variant, we do not partition the input into patches and apply two-direction strip-based self-attention to the full resolution to obtain the global receptive field, as illustrated in Fig. 3.

As represented in Fig. 2(a), different Partition variants are used in the order of Partition-S, Partition-M, and Partition-L in the encoder sub-network, and the opposite order is applied to the decoder. This design has two key advantages. Firstly, we use small patches on high-resolution furthers and large patches for low-resolution features, leading to high efficiency. Secondly, in the encoder, the receptive fields are enlarged gradually, helping encode information in a local-to-global manner. As a result, the deepest features have more information about the locations of degradation blurs. In contrast, in the decoder network, the order of using Partition versions helps the model restore clean features in a coarse-to-fine manner, which is consistent with previous multi-stage frameworks [17].

Complexity analyses: Compared to the strip-based self-attention, whose complexity is $\Omega(\text{Strip}) = HW(H+W)C$, our Partition-S and Partition-M are more efficient with the complexity of $\frac{1}{4}\Omega(\text{Strip})$ and $\frac{1}{2}\Omega(\text{Strip})$, respectively. Partition-L has the same complexity as the strip attention. Note that we do not take into consideration linear layers for simplicity when computing complexity.

3) Overall Pipeline of EAM: With input $I \in \mathbb{R}^{H\times W\times C}$, we first utilize a LayerNorm layer followed by a 1×1 convolution to obtain input features for the Partition part. Then the resulting features are split into two parts among the channel dimension. The above process can be expressed as:

$$I^h, I^v = \text{Split}(f_{1\times 1}(\text{LN}(I))) \tag{5}$$

where $f_{1\times 1}$ and LN denote 1×1 convolution layer and layer normalization, respectively; and I^h, $I^v \in \mathbb{R}^{H\times W\times \frac{C}{2}}$ represent input features of Partition part. After being processed by two-direction attention, the final output of EAM can be obtained by:

$$\hat{I} = f_{1\times 1}([\mathcal{R}(\hat{I}^h), \mathcal{R}(\hat{I}^v)]) + I \tag{6}$$

where $\mathcal{R}$ means reshaping the resulting features of two-direction attention to $H\times W\times \frac{C}{2}$, and $[\cdot,\cdot]$ is concatenation.

C. Loss Function

In this study, we leverage l_1 loss in two domains to train our PSNet. For same-size input/output images, dual-domain loss functions [10] are given by:

$$\mathcal{L}_s = \frac{1}{t}\|\hat{O} - O\|_1, \quad \mathcal{L}_f = \frac{1}{t}\|\mathcal{F}(\hat{O}) - \mathcal{F}(O)\|_1 \tag{7}$$

where t denotes the number of total elements for normalization; $\hat{O}$ and O represent the predicted and ground-truth images, respectively; and $\mathcal{F}$ denotes the fast Fourier transform. The overall loss function is obtained by combining the above two terms: $\mathcal{L} = \mathcal{L}_s + \lambda\mathcal{L}_f$, where λ is set to 0.1.

IV. Experiments

In this part, we first introduce the datasets and implementation details. Next, we present the experimental results. In the tables, the best performance is highlighted in bold. Unless stated otherwise, FLOPs are measured on 256×256 patches.

A. Experimental Setup

1) Datasets: We evaluate PSNet on five restoration tasks: single-image defocus deblurring, image dehazing, image motion deblurring, image desnowing, and image denoising.

Image dehazing: We use the RESIDE [3] dataset for daytime dehazing. It contains two training subsets, i.e., ITS and OTS, and a testing set, SOTS. Following previous methods [5], [18], we evaluate the ITS-trained and OTS-trained models on SOTS-Indoor and SOTS-Outdoor test sets, respectively. In addition to the daytime dataset, we further evaluate our model on the night-time dataset, i.e., NHR [19]. Furthermore, we evaluate PSNet on three real-world datasets: Dense-Haze [20], NH-HAZE [21] and NH-HAZE2 [22]. Models are trained for 300 epochs on ITS [3] and NHR [19], and 30 epochs for OTS [3]. Following [5], the models are trained for 5000 epochs on real-world datasets.

Motion deblurring: We verify the effectiveness of our model on GoPro [4] for motion deblurring. GoPro contains 2103 and 1111 clean/degraded image pairs for training and testing, respectively. Moreover, to demonstrate the generalization ability of our method, we directly apply the GoPro-trained model to the HIDE [23] dataset, which has 2025 image pairs for testing. Following previous methods [10], [17], the model is trained for 3000 epochs on GoPro.

Defocus deblurring: DPDD [24] is leveraged to evaluate the efficacy of PSNet. DPDD contains 500 indoor/outdoor scenes, each with four images labeled as center view, left view, right view, and an all-in-focus ground-truth image. Our model is trained by inputting the center view image. The training strategy follows the previous method [25].

Image desnowing: We adopt three widely used datasets, CSD [26], SRRS [27], and Snow100K [28], for the desnowing task. We train PSNet for 800 epochs on each dataset.

Image denoising: We train PSNet for 120 epochs using the same composite dataset as Restormer [6]. We train a separate model for each noise level.

2) Implementation Details: Adam serves as the optimizer with the initial learning rate as $1e^{-4}$, gradually reduced to $1e^{-6}$ with cosine annealing. We adopt random horizontal flips for data augmentation. With the exception of NHR [19]/denoising and NH-HAZE [21], where the batch size is set to 8 and 2, respectively, four samples are fed into models for each iteration. Models are trained on 256×256 patches except for real-world dehazing datasets where the patch size is 600×800. The number of residual blocks in ResBlock, i.e., n in Fig. 2(b), is set to 3 and 15 for dehazing/desnowing and deblurring/denoising. The number of heads in EAM is set to 4.

B. Experimental Results

1) Image Dehazing: The daytime dehazing results on SOTS-Indoor and SOTS-Outdoor datasets [3] are shown in Table I. Our method achieves better performance than other state-of-the-art algorithms. Specifically, PSNet obtains a significant performance gain of 2.42 dB in terms of PSNR over DeHamer [5] on SOTS-Outdoor with only 3% parameters. In addition, PSNet is superior to DehazeFormer-L [1] by 0.4 dB on SOTS-Indoor with 86% fewer complexity (Fig. 1). We provide visual comparisons

Fig. 4. Image dehazing results on the SOTS-Indoor [3] test set. Zoom in for the best view. The quantitative results are computed on full images.

TABLE I
IMAGE DEHAZING RESULTS ON THE SOTS [3] DATASET

Methods	Outdoor		Indoor		Params	FLOPs
	PSNR	SSIM	PSNR	SSIM	(M)	(G)
GDNet [29]	30.86	0.982	32.16	0.984	0.956	21.49
FFA-Net [12]	33.57	0.984	36.39	0.989	4.456	287.8
DeHamer [5]	35.18	0.986	36.63	0.988	132.45	48.93
PMNet [18]	34.74	0.985	38.41	0.990	18.90	81.13
DehazeFormer-L [1]	-	-	40.05	**0.996**	25.44	279.7
PSNet	**37.60**	**0.995**	**40.45**	**0.996**	4.07	40.08

TABLE II
IMAGE DEHAZING RESULTS ON THE REAL-WORLD DATASETS

Method	Dense-Haze		NH-HAZE		NH-HAZE2		FLOPs/G
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
GDNet [29]	13.31	0.368	13.80	0.537	19.26	0.805	21.49
FFA-Net [12]	14.39	0.452	19.87	0.692	20.00	0.823	287.8
AECR-Net [8]	15.80	0.466	19.88	0.717	20.68	0.828	52.20
DeHamer [5]	16.62	0.560	**20.66**	0.684	-	-	48.93
C^2PNet [30]	16.88	0.573	-	-	21.19	0.833	462.24
PSNet	**16.90**	**0.631**	20.24	**0.796**	**21.51**	**0.894**	40.08

TABLE III
NIGHTTIME IMAGE DEHAZING RESULTS ON THE NHR [19] DATASET

Method	GS [31]	MRPF [32]	MRP [32]	OSFD [19]	HCD [33]	PSNet Ours
PSNR	17.32	16.95	19.93	21.32	23.43	**25.47**
SSIM	0.629	0.667	0.777	0.804	0.953	**0.959**

TABLE IV
IMAGE DESNOWING RESULTS ON THREE WIDELY USED DATASETS: CSD [26], SRRS [27], AND SNOW100K [28]

Method	CSD [26]		SRRS [27]		Snow100K [28]		FLOPs/G
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	
DesnowNet [28]	20.13	0.81	20.38	0.84	30.50	0.94	1.7K
JSTASR [27]	27.96	0.88	25.82	0.89	23.12	0.86	-
SMGARN [34]	31.93	0.95	29.14	0.94	31.92	0.93	450.3
TKL [35]	33.89	0.96	30.82	0.96	**34.37**	0.95	41.58
DGUNet [36]	34.74	0.97	31.28	0.96	34.21	0.95	199.74
PSNet	**37.61**	**0.99**	**31.89**	**0.98**	33.81	**0.96**	40.08

on SOTS-Indoor in Fig. 4. As can be seen, our PSNet is more effective in removing hazy blurs than other methods and produces more visually-pleasing results. In addition, Table II shows that our model obtains favorable results on three real-world datasets. Furthermore, the results on the nighttime dataset NHR [19] are reported in Table III. Our model outperforms the recent algorithm HCD [33] by a large margin of 2.04 dB PSNR. Fig. 6 illustrates that the images yielded by our method are visually closer to the targets.

2) Image Desnowing: The results on three desnowing datasets are reported in Table IV. The proposed PSNet obtains

TABLE V
GAUSSIAN GRAYSCALE IMAGE DENOISING RESULTS ON BSD68 [37]

Method	σ=15	σ=25	σ=50	FLOPs/G
DeamNet [38]	31.91	29.44	26.54	146
DAGL [39]	31.93	29.46	26.51	256
SwinIR [40]	**31.97**	29.50	26.58	759
Restormer [6]	31.96	29.52	26.62	141
PSNet	31.96	**29.53**	**26.65**	127

the best performance in all categories. Specifically, our model outperforms NAFNet [44] on the CSD [26] dataset by a large margin of 4.48 dB in terms of PSRN. In addition, our method is superior to TransWeather [45] on all datasets with only 18.6% parameters. Fig. 5 illustrates that our model produces a more high-quality image than other methods.

3) Image Denoising: Table V shows that our method outperforms Restormer [6] by 0.03 dB (σ=50) and 0.01 dB (σ=25) on the BSD68 dataset [37] with lower FLOPs. The visual comparisons are presented in Fig. 7. The proposed PSNet yields a sharper image than Restormer [6] for σ=50.

4) Single-Image Defocus Deblurring: The results on DPDD [24] are shown in Table VI. Our PSNet receives the best performance on most metrics. Particularly in the indoor scenes category, PSNet significantly outperforms Restormer [6] by 0.43 dB PSNR. Compared to the CNN-based method DRBNet [25], our method produces a substantial gain of 0.63 dB PSNR on the combined scene category. Fig. 8 illustrates that PSNet is more effective in removing spatially-varying degradation blur than other algorithms.

5) Image Motion Deblurring: The results on the GoPro [4] dataset are represented in Table VII. The proposed PSNet obtains a performance gain of 0.07 dB in terms of PSNR over the strong Transformer model Restormer [6] with about half parameters, lower computational complexity, and 3.05× faster speed, as illustrated in Fig. 1 (Right). The results demonstrate the efficacy of our method. The visual results are illustrated in Fig. 9. Our method reconstructs more details from the difficult example than other methods. Moreover, we verify the generalization ability of our framework by directly applying the GoPro-trained model to the HIDE [23] dataset. The results are reported in Table VIII. Our method receives 0.09 dB PSNR performance gain compared to the Transformer-based model Uformer [9].

6) Effects on Downstream Tasks: We study the effects of our method on robotic applications, *e.g.,* detection and segmentation, using YOLOv7 [48] and SAM [49], respectively. Fig. 10 and Fig. 11 show that the deblurring results of PSNet are beneficial for detection and segmentation, respectively. Specifically, the person can be detected in the resulting image of our method.

Fig. 5. Image desnowing results on the CSD [26] dataset. Zoom in for the best view.

TABLE VI
SINGLE-IMAGE DEFOCUS DEBLURRING RESULTS ON THE DPDD [24] DATASET. FLOPS ARE MEASURED ON THE 720×1280 PATCH

Method	Indoor Scenes				Outdoor Scenes				Combined				FLOPs/G
	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	
DPDNet [24]	26.54	0.816	0.031	0.239	22.25	0.682	0.056	0.313	24.34	0.747	0.044	0.277	770
AIFNet [41]			-				-		24.21	0.742	-	0.309	1747
MDP [42]	28.02	0.841	0.027	-	22.82	0.690	0.052	-	25.35	0.763	0.040	-	1898
DRBNet[25]			-				-		25.73	0.791	-	0.183	693
Restormer[6]	28.87	**0.882**	0.025	**0.145**	23.24	0.743	0.050	**0.209**	25.98	0.811	0.038	**0.178**	1983
PSNet	**29.30**	0.880	**0.024**	0.165	**23.57**	**0.753**	**0.049**	0.239	**26.36**	**0.815**	**0.037**	0.203	1790

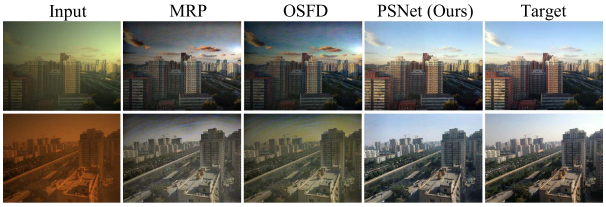


Fig. 6. Nighttime image dehazing results on the NHR [19] dataset.

Fig. 7. Gaussian grayscale denoising comparisons on BSD68 [37].

TABLE VII
IMAGE MOTION DEBLURRING RESULTS ON THE GOPRO [4] DATASET

Method	Date	GoPro [4]		Overhead	
		PSNR	SSIM	Params/M	FLOPs/G
DBGAN [43]	CVPR'20	31.10	0.942	11.6	759.85
MIMO-UNet+ [10]	ICCV'21	32.45	0.957	16.1	154.41
MPRNet [17]	CVPR'21	32.66	0.959	20.1	777.01
MAXIM [16]	CVPR'22	32.86	**0.961**	22.2	169.5
Restormer [6]	CVPR'22	32.92	**0.961**	26.13	140.99
PSNet	Ours	**32.99**	**0.961**	13.37	127.32

Furthermore, Fig. 12 suggests that the produced haze-free image is also useful for the environment perception in bad weather.

C. Ablation Study

We conduct ablation studies to verify the effectiveness of our modules and investigate the influences of different numbers of heads. All experiments are performed on RESIDE-Indoor [3] with $n = 0$ (Fig. 2(b)). Other experimental settings remain identical to that of our final dehazing model (Table I). The baseline is obtained by removing EAM from the tiny PSNet (when $n = 0$).

1) Effects of Each Component: We first study the effects of the proposed Partition methods. The results are shown in

TABLE VIII
IMAGE MOTION DEBLURRING RESULTS ON THE HIDE [23] DATASET

Method	DBGAN [43]	Suin *et al.* [46]	MIMO-UNet+ [10]	HINet [47]	Uformer [9]	PSNet Ours
PSNR	28.94	29.98	29.99	30.32	30.83	**30.92**
SSIM	0.915	0.930	0.930	0.932	**0.952**	0.939

TABLE IX
ABLATION STUDIES FOR DIFFERENT PARTITION METHODS. S/M/L=IA-S/M/L

Variant	Baseline	1st scale	2rd scale	3rd scale	PSNR
a	w/o $\mathcal{L}_f$	-	-	-	30.87
b	✓	-	-	-	31.33
c	✓	S	S	S	34.89
d	✓	S	M	S	35.89
e	✓	S	M	M	35.98
f	✓	S	M	L	**36.20**
g	✓	L	M	S	35.60
h	✓	L	L	L	36.29
i	✓	M	M	M	35.98

Table IX. As can be seen, the baseline model receives 31.33 dB PSNR on the SOTS-Indoor [3] dataset. Without using the frequency loss, the performance degrades to 30.87 dB. Since our main goal of this study is to develop an efficient framework for image restoration, we first employ IA-S in all scales of both encoder and decoder networks, which leads to a performance gain of 3.56 dB PSRN compared to the baseline. By substituting IA-M for IA-S in the second sale of encoder and decoder networks, the model (Table IX d) obtains further performance boost of 1 dB over the variant Table IX c. In addition, the model (Table IX e) with an additional IA-M used in the third scale receives 35.98 dB PSNR, demonstrating the effectiveness of enlarging receptive fields on the deepest features. Our final model (Table IX f) obtains the best performance among methods in the top set by introducing IA-L in the third scale.

Furthermore, we invert the order of using IA modules in Table IX g. The model only receives 35.6 dB, 0.6 dB lower than our choice. In addition, we adopt IA-L in all scales (Table IX h), and the model obtains a performance gain of 0.09 over ours while taking 52% more training time. Deploying IA-M in all scales

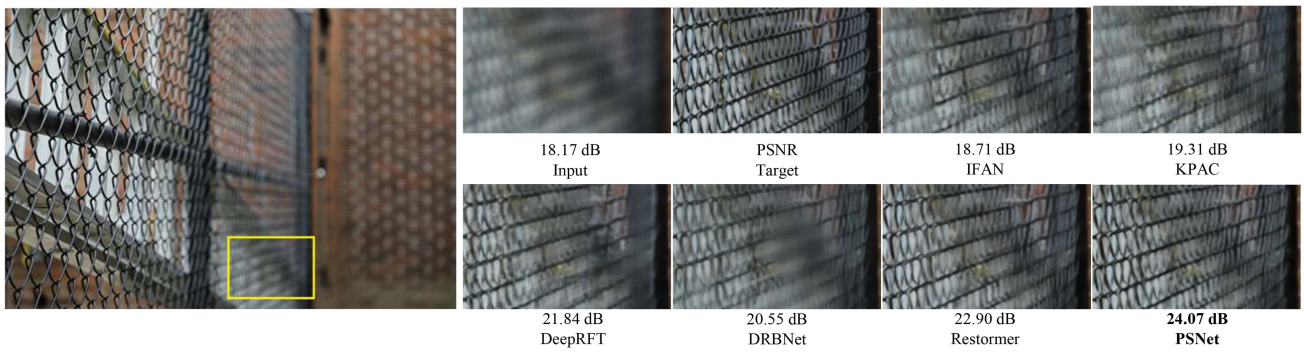


Fig. 8. Single-image defocus deblurring results on the DPDD [24] dataset.

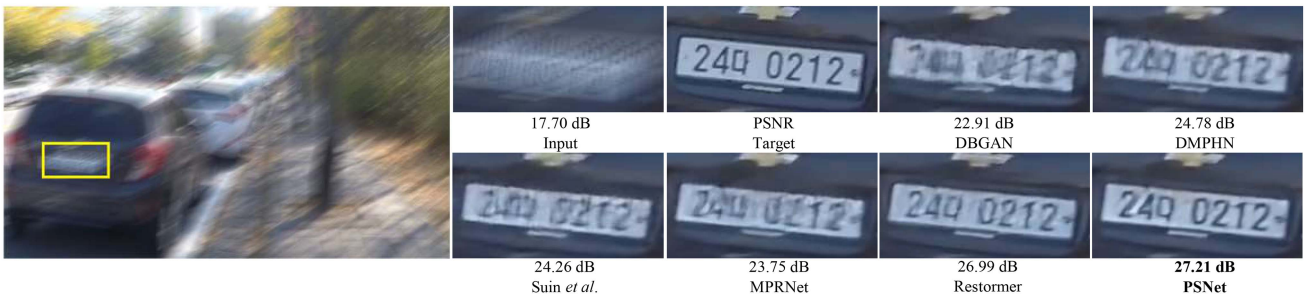
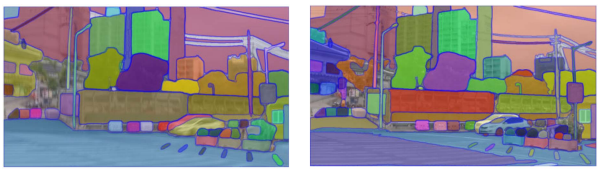


Fig. 9. Image motion deblurring results on the GoPro [4] dataset.

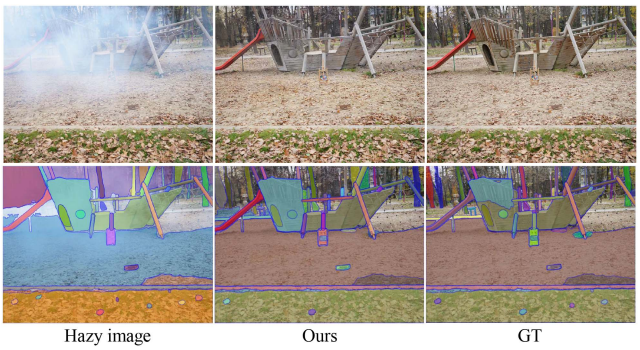
(a) Blurry image (b) Our result

Fig. 10. Detection comparisons between the blurry input and our result. The image is obtained from the GoPro [4] dataset.



(a) Blurry image (b) Our result

Fig. 11. Segmentation comparisons between the blurry input and our result.



Hazy image Ours GT

Fig. 12. Haze-free image yielded by our method is beneficial for segmentation. The image is obtained from NH-HAZE2 [22].

TABLE X
ABLATION STUDIES FOR THE NUMBER OF HEADS IN EAM

Heads	2	4	8	16
PSNR	36.20	**36.40**	36.38	36.30

(Table IX i) receives the same performance as S-M-M (Table IX e), demonstrating the efficacy of our design of applying a small receptive field to high-resolution features. This conclusion can also be drawn by comparing Table IX d and Table IX g, where applying IA-L to large features leads to inferior performance.

2) Number of Heads in EMA: We further investigate the influences of different numbers of heads in EMA by varying d in 2. The results are represented in Table X. The performance improves when increasing the number of heads from 2 to 4 and saturates at 8 heads, which is probably caused by overfitting. Thus, we choose 4 heads in our final model.

V. CONCLUSION

In this study, we propose an efficient framework for image restoration, dubbed PSNet, which elaborately determines the operation region size for self-attention. More specifically, we apply small regions for self-attention on high-resolution features, which improves efficiency and performance simultaneously. In contrast, we impose self-attention on the full-size deepest features, achieving the global receptive field, which is useful to manage large-scale degradation blurs. Besides, the above design also helps the decoder network recovers the clean image in a coarse-to-fine manner. In addition, in each region, we leverage EMA to perform information integration, which is established on the more efficient strip-based self-attention, rather than the

expensive global self-attention. Comprehensive experiments on 13 benchmark datasets demonstrate that our efficient PSNet achieves state-of-the-art performance on five image restoration tasks.

REFERENCES

[1] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Trans. Image Process.*, vol. 32, no. 3, pp. 1927–1941, Mar. 2023.

[2] Y. Cui, Y. Tao, W. Ren, and A. Knoll, "Dual-domain attention for image deblurring.," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 479–487.

[3] B. Li et al., "Benchmarking single image dehazing and beyond," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 492–505, Jan. 2019.

[4] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3883–3891.

[5] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3D position embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5812–5820.

[6] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5728–5739.

[7] S. Chen, T. Ye, Y. Liu, T. Liao, J. Jiang, E. Chen, and P. Chen, "MSP-Former: Multi-scale projection transformer for single image desnowing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2023, pp. 1–5.

[8] H. Wu et al., "Contrastive learning for compact single image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10551–10560.

[9] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17683–17693.

[10] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4641–4650.

[11] Y. Zhang et al., "Kbnet: Kernel basis network for image restoration," 2023, *arXiv:2303.02881*.

[12] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11908–11915.

[13] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2034–2042.

[14] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2022, pp. 146–162.

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

[16] Z. Tu et al., "MaXIM: Multi-axis mlp for image processing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5769–5780.

[17] S. W. Zamir et al., "Multi-stage progressive image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14821–14831.

[18] T. Ye et al., "Perceiving and modeling density for image dehazing," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2022, pp. 130–145.

[19] J. Zhang, Y. Cao, Z.-J. Zha, and D. Tao, "Nighttime dehazing with a synthetic benchmark," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2355–2363.

[20] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images," in *Proc. Int. Conf. Image Process.*, 2019, pp. 1014–1018.

[21] C. O. Ancuti, C. Ancuti, and R. Timofte, "NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 444–445.

[22] C. O. Ancuti, C. Ancuti, F.-A. Vasluianu, and R. Timofte, "NTIRE 2021 nonhomogeneous dehazing challenge report," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 627–646.

[23] Z. Shen et al., "Human-aware motion deblurring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5572–5581.

[24] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2020, pp. 111–126.

[25] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16304–16313.

[26] W.-T. Chen et al., "All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4196–4205.

[27] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "JSTASR: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *Proc. IEEE 16th Eur. Conf. Comput. Vis.*, 2020, pp. 754–770.

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

[29] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 7314–7323.

[30] Y. Zheng, J. Zhan, S. He, J. Dong, and Y. Du, "Curricular contrastive regularization for physics-aware single image dehazing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 5785–5794.

[31] Y. Li, R. T. Tan, and M. S. Brown, "Nighttime haze removal with glow and multiple light colors," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2015, pp. 226–234.

[32] J. Zhang, Y. Cao, S. Fang, Y. Kang, and C. Wen Chen, "Fast haze removal for nighttime image using maximum reflectance prior," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7418–7426.

[33] T. Wang et al., "Restoring vision in hazy weather with hierarchical contrastive learning," 2022, *arXiv:2212.11473*.

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

[35] W.-T. Chen, Z.-K. Huang, C.-C. Tsai, H.-H. Yang, J.-J. Ding, and S.-Y. Kuo, "Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17653–17662.

[36] C. Mou, Q. Wang, and J. Zhang, "Deep generalized unfolding networks for image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17399–17410.

[37] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE/CVF 8th Int. Conf. Comput. Vis.*, 2001, pp. 416–423.

[38] C. Ren, X. He, C. Wang, and Z. Zhao, "Adaptive consistency prior based deep network for image denoising," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 8596–8606.

[39] C. Mou, J. Zhang, and Z. Wu, "Dynamic attentive graph learning for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 4328–4337.

[40] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1833–1844.

[41] L. Ruan, B. Chen, J. Li, and M.-L. Lam, "AIFNet: All-in-focus image restoration network using a light field-based dataset," *IEEE Trans Comput Imag.*, vol. 7, no. 5, pp. 675–688, Jun. 2021.

[42] A. Abuolaim, M. Afifi, and M. S. Brown, "Improving single-image defocus deblurring: How dual-pixel images help through multi-task learning," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1231–1239.

[43] K. Zhang et al., "Deblurring by realistic blurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2737–2746.

[44] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2022, pp. 17–33.

[45] J. M. J. Valanarasu, R. Yasarla, and V. M. Patel, "Transweather: Transformer-based restoration of images degraded by adverse weather conditions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2353–2363.

[46] M. Suin, K. Purohit, and A. N. Rajagopalan, "Spatially-attentive patch-hierarchical network for adaptive motion deblurring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3606–3615.

[47] L. Chen, X. Lu, J. Zhang, X. Chu, and C. Chen, "Hinet: Half instance normalization network for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 182–192.

[48] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7464–]7475.

[49] A. Kirillov et al., "Segment anything," 2023, *arXiv:2304.02643*.