DiNAT-IR: Exploring Dilated Neighborhood Attention for High-Quality Image Restoration

Anonymous authors
Paper under double-blind review

Abstract

Transformers, with their self-attention mechanisms for modeling long-range dependencies, have become a dominant paradigm in image restoration tasks. However, the high computational cost of self-attention limits scalability to high-resolution images, making efficiency–quality trade-offs a key research focus. To address this, Restormer employs channel-wise self-attention, which computes attention across channels instead of spatial dimensions. While effective, this approach may overlook localized artifacts that are crucial for high-quality image restoration. To bridge this gap, we explore Dilated Neighborhood Attention (DiNA) as a promising alternative, inspired by its success in high-level vision tasks. DiNA balances global context and local precision by integrating sliding-window attention with mixed dilation factors, effectively expanding the receptive field without excessive overhead. However, our preliminary experiments indicate that directly applying this global-local design to the classic deblurring task hinders accurate visual restoration, primarily due to the constrained global context understanding within local attention. To address this, we introduces a channel-aware module that complements local attention, effectively integrating global context without sacrificing pixel-level precision. The proposed DiNAT-IR, a Transformer-based architecture specifically designed for image restoration, achieves competitive results across multiple benchmarks, offering a high-quality solution for diverse low-level computer vision problems. Our codes will be released soon.

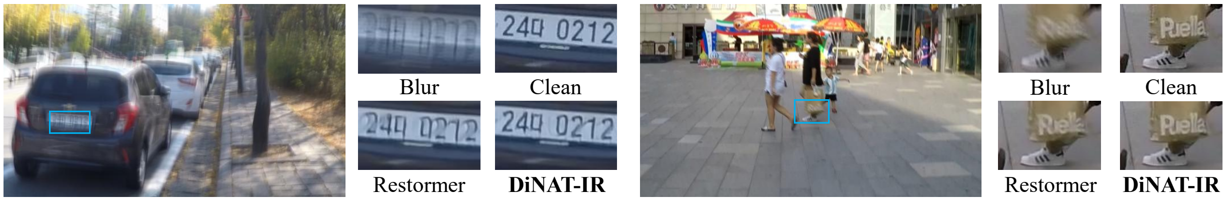
1 Introduction

Image restoration is a fundamental task in computer vision, with wide-ranging applications in fields such as autonomous driving, medical imaging, and satellite remote sensing (Ding et al., 2021; Zhang & Dong, 2020; Rasti et al., 2021). It aims to recover high-quality images from degraded inputs, addressing challenges like blur, noise, and the other types of artifacts (Banham & Katsaggelos, 1997).

In recent years, Transformers have emerged as powerful models for image restoration. Unlike traditional convolutional neural networks (CNNs) that rely on staked convolutional layers (Zhang et al., 2017b; Zamir et al., 2021; Chen et al., 2022), Transformers utilize self-attention to model long-range pixel relationships (Liang et al., 2021; Wang et al., 2022; Zamir et al., 2022), making them particularly effective for low-level computer vision tasks like deblurring, denoising, deraining, and super-resolution.

Despite their effectiveness, balancing the computational cost of self-attention with restoration quality remains a key challenge, especially for high-resolution images. Restormer (Zamir et al., 2022) addresses this by computing self-attention along the channel dimension instead of the spatial domain, achieving a strong trade-off between efficiency and performance. However, recent studies report that this design misses local details, as shown in Figure 1, which are critical in dynamic scenes (Jang et al., 2023; Chen et al., 2024).

To bridge this gap, we explore Dilated Neighborhood Attention (DiNA) as a promising alternative, inspired by its recent success in detection and segmentation (Hassani & Shi, 2022). Unlike previous self-attention mechanisms, which either aggregate global context entirely or focus solely on local patches, DiNA integrates sliding-window attention with mixed dilation factors, effectively expanding the receptive field without incurring excessive computational overhead. The original DiNAT (Hua et al., 2019) reports that a hybrid



(a) GoPro Dataset

(b) HIDE Dataset

Figure 1: Visual comparisons between Restormer (Zamir et al., 2022) and our proposed DiNAT-IR on the motion deblurring datasets (Nah et al., 2017; Shen et al., 2019). DiNAT-IR produces cleaner restoration of numbers and characters on car license plates and hand-held bags. Zoom in to see details.

design, using local neighborhood attention (NA) with a dilation factor $\delta = 1$ alongside global DiNA, improves performance in high-level computer vision tasks.. However, our preliminary experiments reveal that directly applying this hybrid design to motion deblurring results in a notable performance drop compared to global-DiNA-only methods. We attribute this to the limited global context understanding of local NA, which restricts its ability to recover clean structures in full-resolution images.

To address this challenge, we introduce a channel-aware module that complements local attention by efficiently integrating global context without sacrificing pixel-level precision. This design effectively addresses the aforementioned bottleneck, allowing for more comprehensive feature interactions across the entire image. Furthermore, the proposed architecture, DiNAT-IR, has achieved competitive results across multiple benchmarks, demonstrating its potential as a high-fidelity solution for diverse image restoration challenges.

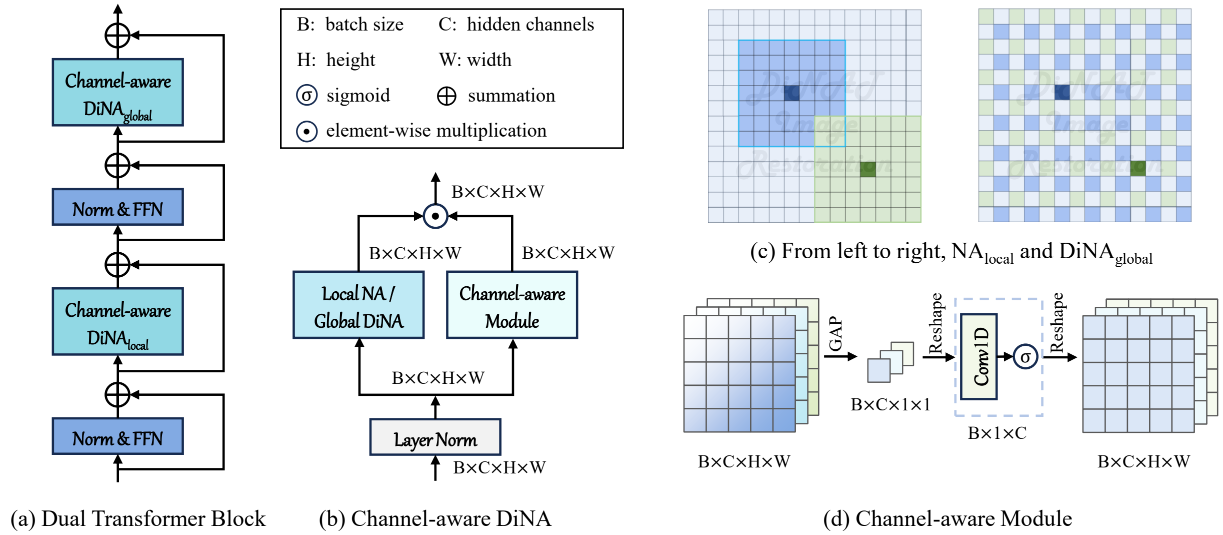
Our main contributions are threefold:

- We investigate the application of dilated neighborhood attention for image deblurring and identify key limitations of its hybrid attention design in this context.
- We introduce a simple while effective channel-aware module that complements local neighborhood attention and restores global context without sacrificing pixel-level detail.
- We propose DiNAT-IR, a Transformer-based architecture that achieves competitive performance not only on deblurring benchmarks but also on other restoration tasks.

2 Related Work

CNNs for Image Restoration. Convolutional neural networks (CNNs) consistently demonstrate strong performance across low-level computer vision tasks. DnCNN (Zhang et al., 2017b) pioneers the use of residual learning for image denoising, laying the groundwork for deeper and more effective architectures. MPRNet (Zamir et al., 2021) adopts a multi-stage framework that processes image features at multiple spatial scales, achieving state-of-the-art results in image restoration. In the era of Transformer-based models, NAFNet (Chen et al., 2022) stands out by showing that, with proper optimization, compact and purely convolutional architectures can still rival more complex Transformer designs in both efficiency and performance. Nevertheless, a key limitation of CNN-based approaches lies in their reliance on deeply stacked convolutional layers to enlarge the receptive field, which restricts their ability to model long-range dependencies effectively.

Transformers for Image Restoration. In contrast, Transformer-based architectures inherently model global context through self-attention mechanisms. While applying vanilla Transformers (Vaswani, 2017) to high-resolution images faced challenges due to the quadratic computational complexity of self-attention with respect to spatial dimensions, subsequent architectural innovations have significantly mitigated this issue in low-level computer vision tasks. For example, SwinIR (Liang et al., 2021) combines convolutional layers for shallow feature extraction with shifted window-based Transformer blocks to capture deeper representations, achieving strong performance in tasks such as super-resolution and denoising. Uformer (Wang et al., 2022) integrates Locally-enhanced Window (LeWin) attention within a U-Net structure, effectively preserving spatial detail for deblurring and deraining tasks. Different from window-based methods, Restormer (Zamir et al.,



(c) From left to right, NA$_{local}$ and DiNA$_{global}$

(a) Dual Transformer Block

(b) Channel-aware DiNA

(d) Channel-aware Module

Figure 2: Structures of (a) the Dual Transformer block with the alternating NA-DiNA attention scheme, (b) the channel-aware DiNA module, (c) local DiNA (NA) and global DiNA (DiNA) blocks, and (d) the channel-aware module. Note: GAP denotes global average pooling, and Conv1D indicates 1D convolution.

2022) improves computational efficiency by computing self-attention along the channel dimension rather than spatial dimensions. However, follow-up studies (Jang et al., 2023; Chen et al., 2024) observe that such designs may overlook fine-grained local details that are critical for restoration in real-world environments.

Dilated Neighborhood Attention. Recent advancements in vision Transformers have prioritized improving the efficiency of self-attention mechanisms while preserving their ability to capture long-range dependencies. Hierarchical models such as the Swin Transformer (Liu et al., 2021) and the Neighborhood Attention Transformer (Hassani et al., 2023) reduce computational costs by restricting self-attention to local windows. However, this often comes at the expense of the global receptive field, an essential attribute for high-level visual understanding. To address this limitation, *Dilated Neighborhood Attention* (**DiNA**) (Hassani & Shi, 2022) extends *neighborhood attention* (**NA**) by sparsifying it across dilated local regions. This design enables an exponential increase in the receptive field without incurring additional computational overhead. Their resulting model, the Dilated Neighborhood Attention Transformer, with dense local **NA** and sparse global **DiNA** (abbreviated as **NA-DiNA**), achieving strong performance in high-level computer vision tasks such as object detection, instance segmentation, and semantic segmentation. Despite these strengths, we observe that directly applying the original **NA-DiNA** method to low-level computer vision tasks like motion deblurring results in a noticeable performance drop compared to the **DiNA-only** attention design. This may be attributed to the inherently limited global context understanding of local **NA**, which struggles to fully capture the spatial extent and complexity of image degradation patterns common in motion blur scenarios.

3 Method

The overall pipeline of DiNAT-IR is based on Restormer (Zamir et al., 2022). It adopts a multi-level U-Net structure that efficiently captures degradation patterns through hierarchical feature processing. The encoder gradually downsamples the input to extract deep features, while the decoder upsamples and refines the output using skip connections that preserve spatial resolution. We build upon this framework and integrate an improved attention mechanism, which is detailed in the following sections.

3.1 Alternating NA-DiNA Attention Scheme

To effectively model both fine-grained structures and large-scale degradation patterns, DiNAT-IR integrates an alternating NA-DiNA strategy within its Transformer blocks, drawing inspiration from the Dilated Neighborhood Attention Transformer (DiNAT) (Hassani & Shi, 2022). By setting the dilation factor δ to 1, DiNA effectively reduces to standard Neighborhood Attention (NA) (Hassani et al., 2023). At each level of DiNAT-IR, the self-attention blocks alternate between two dilation factors to vary the attention window size. Specifically, setting the dilation factor $\delta = 1$ yields local NA, while larger values of δ expand the receptive field to capture broader context. The dilation pairs are defined as $\delta \in \{1, 36\}$, $\{1, 18\}$, $\{1, 9\}$, and $\{1, 4\}$ across the four stages of the network, corresponding to progressively finer spatial resolutions. This alternating pattern allows DiNAT-IR to adaptively integrate both local details and global contextual information, improving its capacity to model spatially extensive degradations without introducing significant computational overhead.

While the original **NA-DiNA** architecture was developed for high-level vision tasks, its hybrid attention design can also be intuitively extended to image restoration problems. In this context, the local **NA** is expected to model short-range, pixel-level dependencies, while the sparse **DiNA** captures broader degradation patterns. However, our preliminary experiments reveal that directly applying the vanilla **NA-DiNA** configuration to the low-level tasks such as motion deblurring leads to a noticeable performance drop compared to a **DiNA-only** baseline. We attribute it to the significantly reduced global context understanding introduced by the frequent use of local **NA**. To address this limitation, we propose a lightweight channel-aware module designed to preserve global context modeling while mitigating the drawbacks of overly localized attention.

3.2 Channel Aware Self Attention

Figure 2 (a) shows that channel-aware self-attention contains two parallel units, the self-attention layers (SA) and a channel-aware module (CAM). DiNAT-IR uses alternating neighborhood attention (**NA**) and dilated neighborhood attention (**DiNA**) as the basic component of SA. Furthermore, CAM is proposed to solve the issue of limited receptive filed caused by **NA**. As illustrated in Figure 2 (c), a CAM first transforms the normalized 2-D features into 1-D data by global average pooling (GAP) (Lin et al., 2013; Chu et al., 2022); then, it applies a 1-D convolution to the intermediate features along the channel dimension; finally, a *sigmoid* function is adopted to compute attention scores. The outputs of the CAM and DiNA are merged by element-wise multiplications.

Given a layer normalized input tensor $\mathbf{X} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$. The output of CASA is computed as,

$$\begin{aligned}\hat{\mathbf{X}} &= \mathrm{SA}(\mathbf{X}) \odot \mathrm{CAM}(\mathbf{X}),\\ \mathrm{CAM}(\mathbf{X}) &= f^{-1}(\mathrm{Conv}_{1d}(f(\mathrm{GAP}_{2d}^{1\times1}(\mathbf{X})))),\end{aligned} \tag{1}$$

where $\odot$ denotes element-wise multiplication; f is a tensor manipulation function which squeezes and transposes a $C \times 1 \times 1$ matrix, resulting in a $1 \times C$ matrix; Conv_{1d} denotes a 1D convolution with a kernel size of 3; $\mathrm{GAP}_{2d}^{1\times1}$ indicates global average pooling, outputting a tensor of size 1×1. We employed the GAP design proposed by Chu et al., and the CAM idea draws inspiration from ECA-Net (Wang et al., 2020).

4 Experiments and Analysis

We evaluate the performance of DiNAT-IR across four distinct image restoration tasks: (a) single-image motion deblurring, (b) defocus deblurring with dual inputs and single images, (c) single image deraining, and (d) single image denoising. In the result tables, the best-performing and second-best methods are indicated using **bold** and underline formatting respectively. We primarily compare against multi-task image restoration networks, supplemented by task-specific methods for completeness.

Implementation Details. DiNAT-IR adopts the four-stage U-Net architecture of Restormer (Zamir et al., 2022) as its backbone. All experiments are conducted using a batch size of 16 across 8 NVIDIA A100 GPUs. Task-specific training configurations vary depending on the particular restoration task and dataset.

GoPro (Nah et al., 2017). For the motion deblurring task, we train DiNAT-IR with image patches of size 256 × 256 and a batch size of 16 for 600K iterations using PSNR loss. The initial learning rate is set to 3×10^{-4} and gradually reduced to 1×10^{-6} following a cosine annealing schedule. We use AdamW as the optimizer with betas set to $[0.9, 0.999]$ (Loshchilov & Hutter, 2017). We further fine-tune the network with an image size of 384×384 and a batch size of 8 for an additional 200K iterations, inspired by the progressive training strategy employed in Restormer (Zamir et al., 2022) and Stripformer (Tsai et al., 2022). During fine-tuning, the initial learning rate is set to 1×10^{-4}. We observe that DiNAT-IR may not fully converge to an optimal solution, suggesting that improved training strategies could further enhance performance on the GoPro dataset. The final model used for evaluation is obtained from the last training iteration.

DPDD (Abuolaim & Brown, 2020). For the dual-pixel defocus deblurring task, the dual-input variant of DiNAT-IR is trained with an image size of 256 × 256 and a batch size of 16 for 300K iterations using PSNR loss. The optimizer and learning rate schedule are consistent with those used for motion deblurring. The model at the 290K iteration is selected as our dual-pixel defocus deblurring model. For the single-image defocus deblurring task, we re-train DiNAT-IR with single images as inputs and adopt the model checkpoint at the 140K iteration as the final version.

Rain13K (Jiang et al., 2020). For the single image deraining task, DiNAT-IR is trained with an image size of 256 × 256 and a batch size of 16 for 300K iterations using L1 loss. The optimizer and learning rate schedule follow the same settings as in motion deblurring. We further fine-tune the network with an image size of 384 × 384 and a batch size of 8 for an additional 100K iterations, selecting the model at the 40K fine-tuning iteration for our deraining experiments.

SIDD (Abdelhamed et al., 2018). For the real-world image denoising task, DiNAT-IR is trained with an image size of 256 × 256 and a batch size of 16 for 300K iterations using PSNR loss. The optimizer and learning rate schedule are identical to those in the motion deblurring task. We choose the model at the 220K iteration as our final denoising model.

Table 1: Comparisons of image restoration models on GoPro (Nah et al., 2017) and HIDE (Shen et al., 2019) datasets. We follow MaIR (Li et al., 2025) and report PSNR, SSIM, Params (M), and FLOPs (G).The proposed DiNAT-IR has achieved competitive performance compared to recent restoration networks.

Method	GoPro		HIDE		Model Complexity	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	Params (M) ↓	FLOPs (G) ↓
SRN 2018	30.26	0.934	28.36	0.904	3.76	35.87
DBGAN 2020	31.10	-	28.94	-	11.59	379.92
MT-RNN 2020	31.15	-	29.15	-	**2.64**	**13.72**
DMPHN 2022	31.20	-	29.09	-	86.80	-
CODE 2023	31.94	-	29.67	-	12.18	22.52
MIMO 2021	32.45	0.956	29.99	0.930	16.10	38.64
MPRNet 2021	32.66	0.958	30.96	0.939	20.13	194.42
Restormer 2022	32.92	0.961	31.22	0.942	26.13	35.31
Uformer 2022	33.06	0.967	30.90	**0.953**	50.88	22.36
CU-Mamba 2024	33.53	-	31.47	-	19.70	-
NAFNet 2022	33.69	0.966	31.32	0.942	67.89	15.85
MaIR 2025	33.69	**0.969**	31.57	0.946	26.29	49.29
DiNAT-IR	**33.80**	0.967	**31.57**	0.945	25.90	45.62

4.1 Motion Deblurring Results

We conduct a thorough evaluation of various image restoration models on the GoPro (Nah et al., 2017) and HIDE (Shen et al., 2019)) datasets. As summarized in Table 1, our proposed DiNAT-IR consistently achieves strong results across the board. Specifically, it reaches a PSNR of 33.80 dB on GoPro and 31.57 dB on HIDE, matching or surpassing all compared methods, including the recent high-performing MaIR (Li et al., 2025) and NAFNet(Chen et al., 2022). While MaIR reports a comparable PSNR on both datasets, DiNAT-IR achieves this with slightly fewer parameters and competitive FLOPs. Compared to traditional convolution-based models like MPRNet (Zamir et al., 2021) and attention-based method Restormer (Zamir et al., 2022), DiNAT-IR maintains similar or superior accuracy while preserving efficiency. Furthermore, despite being trained solely on GoPro, DiNAT-IR demonstrates excellent generalization to HIDE, under-

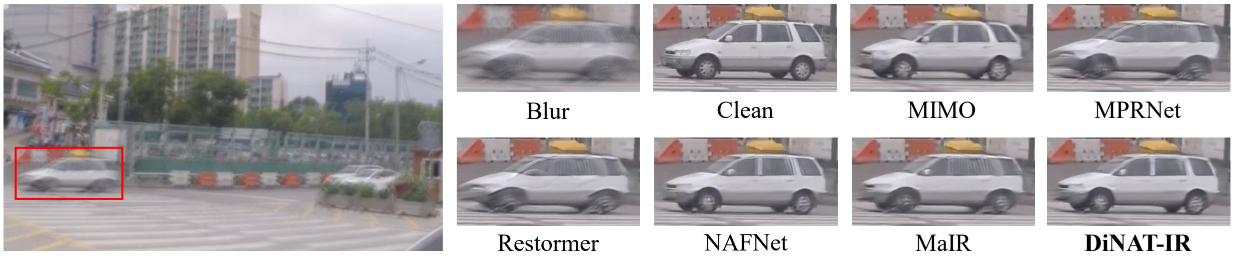


Figure 3: Deblurring results on the GoPro dataset (Nah et al., 2017). Zoom in to see details.

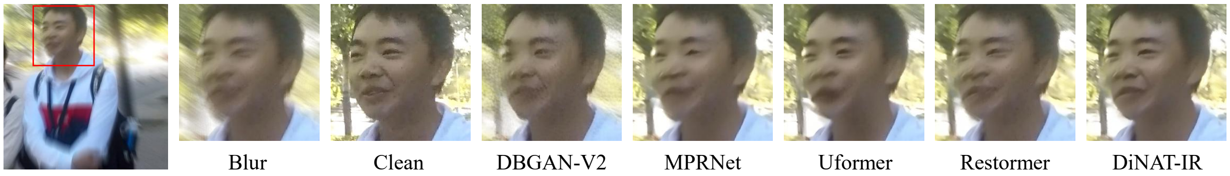


Figure 4: Deblurring results on the HIDE dataset (Shen et al., 2019). Zoom in to see details.

scoring its robustness in human-centric scenarios. These results highlight DiNAT-IR's effective trade-off between model complexity and restoration quality, as well as its potential as a strong alternative in dynamic scenes. Qualitative comparisons in Figure 3 and Figure 4 clearly demonstrate that the image deblurred by our method is more visually closer to the ground-truth than those of the other algorithms.

4.2 Defocus Deblurring Results

Table 2 presents a comprehensive comparison of single-image and dual-pixel defocus deblurring methods on the DPDD dataset (Abuolaim & Brown, 2020). For single-image defocus deblurring, DiNAT-IR$_S$ delivers competitive results, achieving strong performance across all metrics. It obtains the second-highest PSNR and MAE on outdoor scenes and ranks closely behind GRL-B$_S$ (Li et al., 2023) and CSformer$_S$ (Duan et al., 2023) overall. Notably, while GRL-B$_S$ slightly surpasses DiNAT-IR$_S$ in combined PSNR (26.18 dB vs. 26.14 dB), DiNAT-IR$_S$ demonstrates comparable or better PSNR and SSIM on the outdoor scene.

In the dual-pixel setting, DiNAT-IR$_D$ shows excellent performance, either outperforming or closely matching state-of-the-art methods. It achieves the highest PSNR on outdoor scenes (24.47 dB) and the best combined PSNR (27.05 dB), while maintaining competitive SSIM and lowest MAE scores. Compared to Restormer$_D$, which performs strongly indoors, DiNAT-IR$_D$ offers superior outdoor performance and better balance across scenes. These results highlight DiNAT-IR's capability to handle both single-image and dual-pixel defocus deblurring tasks effectively, achieving state-of-the-art performance on the DPDD benchmark.

4.3 Deraining Results

Table 3 summarizes the performance of several image deraining models across five benchmark datasets. DiNAT-IR demonstrates excellent results, achieving SSIM scores nearly identical to those of Restormer (Zamir et al., 2022) across all five datasets, indicating strong perceptual quality and effective detail preservation. Although DiNAT-IR's PSNR is slightly lower than Restormer's, the differences are minor, for example, on the Rain100L test set, DiNAT-IR attains 38.93 dB compared to Restormer's 38.99 dB, a negligible gap considering the task complexity. Compared to earlier methods such as SEMI (Wei et al., 2019), DIDMDN (Zhang & Patel, 2018), and UMRL (Yasarla & Patel, 2019), DiNAT-IR delivers significant improvements in both PSNR and SSIM. It also performs competitively against recent models like MPRNet (Zamir et al., 2021) and SPAIR (Purohit et al., 2021), surpassing them in several metrics. Overall, these results highlight DiNAT-IR as a highly effective deraining model, delivering competitive perceptual quality and achieving performance close to that of Restormer in pixel-level restoration accuracy.

Table 2: **Dual-Pixel Defocus Deblurring comparisons** on the DPDD dataset (Abuolaim & Brown, 2020), which includes 37 indoor and 39 outdoor scenes. D indicates network variants using dual-image inputs; S denotes the single-image task. DiNAT-IR demonstrates performance comparable to GRL-B (Li et al., 2023) across both single-image and dual-pixel settings.

Method	Indoor Scenes			Outdoor Scenes			Combined		
	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓
EBDB$_S$ 2017	25.77	0.772	0.040	21.25	0.599	0.058	23.45	0.683	0.049
DMENet$_S$ 2019	25.50	0.788	0.038	21.43	0.644	0.063	23.41	0.714	0.051
JNB$_S$ 2015	26.73	0.828	0.031	21.10	0.608	0.064	23.84	0.715	0.048
DPDNet$_S$ 2020	26.54	0.816	0.031	22.25	0.682	0.056	24.34	0.747	0.044
KPAC$_S$ 2021	27.97	0.852	0.026	22.62	0.701	0.053	25.22	0.774	0.040
IFAN$_S$ 2021	28.11	0.861	0.026	22.76	0.720	0.052	25.37	0.789	0.039
Restormer$_S$ 2022	28.87	0.882	0.025	23.24	0.743	0.050	25.98	0.811	0.038
CSformer$_S$ 2023	29.01	0.883	**0.023**	**23.63**	0.759	**0.047**	**26.25**	0.819	**0.036**
GRL-B$_S$ 2023	**29.06**	**0.886**	0.024	23.45	**0.761**	0.049	26.18	**0.822**	0.037
DiNAT-IR$_S$	28.94	0.881	0.025	23.48	0.751	0.049	26.14	0.814	0.037
DPDNet$_D$ 2020	27.48	0.849	0.029	22.90	0.726	0.052	25.13	0.786	0.041
RDPD$_D$ 2021	28.10	0.843	0.027	22.82	0.704	0.053	25.39	0.772	0.040
Uformer$_D$ 2022	28.23	0.860	0.026	23.10	0.728	0.051	25.65	0.795	0.039
IFAN$_D$ 2021	28.66	0.868	0.025	23.46	0.743	0.049	25.99	0.804	0.037
Restormer$_D$ 2022	29.48	0.895	0.023	23.97	0.773	0.047	26.66	0.833	0.035
CSformer$_D$ 2023	29.54	0.896	0.023	24.38	0.788	0.045	26.89	0.841	0.034
GRL-B$_D$ 2023	**29.83**	**0.903**	**0.022**	24.39	0.795	0.045	27.04	**0.847**	0.034
DiNAT-IR$_D$	29.76	0.901	0.023	**24.47**	**0.795**	**0.045**	**27.05**	0.846	**0.034**

Table 3: **Image deraining results.** DiNAT-IR achieves performance very close to that of Restormer (Zamir et al., 2022), with SSIM scores nearly matching those of Restormer across multiple datasets. However, we acknowledge noticeably lower PSNR scores for DiNAT-IR on these datasets.

Method	Rain100H		Rain100L		Test2800		Test1200		Test100	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
SEMI 2019	16.56	0.486	25.03	0.842	24.43	0.782	26.05	0.822	22.35	0.788
DIDMDN 2018	17.35	0.524	25.23	0.741	28.13	0.867	29.65	0.901	22.56	0.818
UMRL 2019	26.01	0.832	29.18	0.923	29.97	0.905	30.55	0.910	24.41	0.829
RESCAN 2018	26.36	0.786	29.80	0.881	31.29	0.904	30.51	0.882	25.00	0.835
PreNet 2019	26.77	0.858	32.44	0.950	31.75	0.916	31.36	0.911	24.81	0.851
MSPFN 2020	28.66	0.860	32.40	0.933	32.82	0.930	32.39	0.916	27.50	0.876
MPRNet 2021	30.41	0.890	36.40	0.965	33.64	0.938	32.91	0.916	30.27	0.897
SPAIR 2021	30.95	0.892	36.93	0.969	33.34	0.936	33.04	0.922	30.35	0.909
Restormer 2022	**31.46**	**0.904**	**38.99**	**0.978**	**34.18**	**0.944**	**33.19**	**0.926**	**32.00**	**0.923**
DiNAT-IR	31.26	0.903	38.93	0.977	33.91	0.943	32.31	0.923	31.22	0.920

4.4 Denoising Results

Table 4 compares several real-image denoising methods based on PSNR and SSIM metrics. Early approaches like DnCNN (Zhang et al., 2017a) and BM3D (Dabov et al., 2007) achieve substantially lower performance, with PSNRs below 26 dB and SSIM under 0.70, reflecting limited effectiveness on challenging real-world noise. Modern deep networks such as VDN (Yue et al., 2019), MIRNet (Zamir et al., 2020), MPRNet (Zamir et al., 2021), DAGL (Mou et al., 2021), and Uformer (Wang et al., 2022) demonstrate significant improvements, achieving PSNR values around 39–40 dB and SSIM above 0.95, highlighting the advances brought by learning-

Table 4: **Real image denoising results.** All methods are trained and tested on the SIDD dataset (Abdelhamed et al., 2018). DiNAT-IR achieves performance comparable to Restormer (Zamir et al., 2022).

Method	DnCNN 2017a	BM3D 2007	VDN 2019	MIRNet 2020	MPRNet 2021	DAGL 2021	Uformer 2022	Restormer 2022	MambaIR 2024	**DiNAT-IR** (Ours)
PSNR ↑	23.66	25.65	39.28	39.72	39.71	38.94	39.89	**40.02**	39.89	39.89
SSIM ↑	0.583	0.685	0.956	0.959	0.958	0.953	0.960	0.960	0.960	**0.960**

Table 5: Ablation study on dilation factor configurations and the proposed channel-aware self-attention on the GoPro (Nah et al., 2017) dataset. The baseline is Restormer (Zamir et al., 2022) with 16 hidden channels. **NA** denotes local neighborhood attention while **DiNA** represents sparse dilated neighborhood attention; with and without are abbreviated as w/ and w/o respectively; **CAM** is the proposed channel-aware module. The adopted NA-DiNA with CAM method shows the strongest or competitive quantitative visual results as rated by both distortion and perception metrics.

Networks	**Distortion**		**Perception**		**Params**	**MACs**
	PSNR ↑	SSIM ↑	FID ↓	LPIPS ↓	(M) ↓	(G) ↓
Restormer (baseline)	30.32	0.934	15.16	0.137	3.0	17.30
NA w/o CAM	31.56	0.948	11.12	0.111	3.0	**16.64**
NA w/ CAM	31.97	0.952	10.86	0.105	3.0	16.71
DiNA w/o CAM	32.03	0.953	10.17	0.103	3.0	**16.64**
DiNA w/ CAM	32.06	0.952	10.14	0.103	3.0	16.71
NA-DiNA w/o CAM	31.87	0.951	10.08	0.107	3.0	**16.64**
NA-DiNA w/ CAM	**32.06**	**0.953**	**9.53**	**0.103**	3.0	16.71

based architectures. Among these methods, Restormer (Zamir et al., 2022) achieves the highest performance, with a PSNR of 40.02 dB and an SSIM of 0.960, establishing a strong benchmark. MambaIR (Guo et al., 2024) achieves a PSNR of 39.89 dB and an SSIM of 0.960, closely following Restormer. Our method, DiNAT-IR, also achieves highly competitive results, matching the highest SSIM score of 0.960 and attaining a PSNR of 39.89 dB. Overall, DiNAT-IR performs at the same level as MambaIR, demonstrating strong capabilities in preserving fine details and delivering perceptually pleasing restorations in real-world scenarios.

5 Ablation Study

In this section, we use Restormer (Zamir et al., 2022) with 16 hidden channels as the baseline model. We maintain the overall architecture, including the total number of Transformer blocks, feed-forward networks (FFNs), and feature fusion strategy, as well as consistent training settings on 4 NVIDIA A100 GPUs. All networks are trained and evaluated on the GoPro dataset (Nah et al., 2017), chosen for its ability to ensure stable training across models. We assess restoration quality using both distortion-based metrics (PSNR and SSIM) and perception-based metrics, FID (Heusel et al., 2017; Parmar et al., 2022), LPIPS (Zhang et al., 2018) and NIQE, for comprehensive comparisons. Additionally, we report the total number of parameters and MACs to indicate model complexity.

As shown in Table 5, the local-attention-only network already outperforms the Restormer baseline (Zamir et al., 2022) by 1.24 dB in PSNR, despite being the weakest among the proposed configurations. Introducing our channel-aware module further improves the local variant by 0.41 dB, and also enhances the global-only and hybrid variants by 0.02 dB and 0.19 dB, respectively. The hybrid configuration with the channel-aware module achieves the best overall performance across all metrics. These results validate our finding that the original DiNA design (Hassani & Shi, 2022) with hybrid-attention is suboptimal for deblurring, and demonstrate that the proposed channel-aware module effectively addresses this limitation. Moreover, DiNAT-IR retains a similar parameter count while reducing MACs by 0.59G compared to the baseline model.

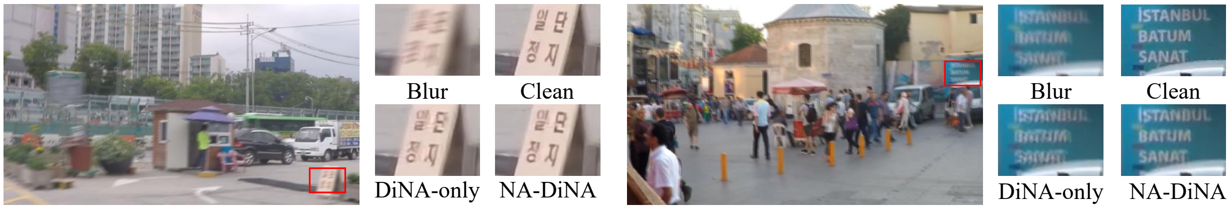


Figure 5: Visual comparisons between DiNAT-IR with global dilated neighborhood attention (DiNA) only and DiNAT-IR with both global DiNA and local neighborhood attention (NA). Both networks are trained and tested on the GoPro dataset (Nah et al., 2017) with the same training settings.

Overall, it achieves a notable 1.74 dB PSNR improvement over the Restormer baseline, offering a superior trade-off between performance and efficiency.

Importantly, although Table 5 shows that the PSNR, SSIM and LPIPS differences between DiNA with CAM and NA-DiNA with CAM are minimal, our observations reveal that incorporating local neighborhood attention (NA) improves the visual quality of the restored images. As illustrated in Figure 5, the method relying solely on global dilated neighborhood attention produces distorted text on the board, whereas including NA results in sharper and more accurate restoration. Therefore, we select DiNAT-IR with NA-DiNA and CAM as our final architecture for the image restoration tasks.

6 Conclusion

In this work, we propose DiNAT-IR, a novel Transformer-based architecture for image restoration that effectively balances global context modeling and local detail preservation. Building upon the strengths of Dilated Neighborhood Attention (DiNA), DiNAT-IR introduces a channel-aware module that enhances global context integration while maintaining pixel-level precision of local Neighborhood Attention (NA). Our experiments demonstrate that, although DiNAT-IR does not consistently surpass Restormer (Zamir et al., 2022) in all metrics, for instance, achieving slightly lower PSNR on certain deraining benchmarks, it delivers comparable or superior restoration performance, particularly in challenging tasks like motion deblurring and defocus deblurring. Furthermore, DiNAT-IR achieves this high-quality restoration with similar or reduced computational costs, offering a favorable trade-off between restoration quality and efficiency. These findings highlight DiNAT-IR as a promising and versatile solution for diverse low-level vision tasks.

7 Limitation

Our ablation studies were conducted primarily on the GoPro dataset (Nah et al., 2017) to ensure consistent training strategies across all models and enable fair comparison. However, this might limit the generalizability of our results to other datasets. Extending the analysis to broader tasks is non-trivial, as it requires substantial effort to adapt training strategies to different data distributions. Since architectural components may yield varied gains across tasks and dataset bias exists, a universally optimal design remains challenging. Future work will explore task-specific architectures to improve generalization and robustness.

Broader Impact Statement

The proposed DiNAT-IR framework advances the field of image restoration by improving the quality and efficiency of deblurring, deraining, and denoising tasks. Positive societal impacts include potential applications in photography, surveillance, autonomous driving, medical imaging, and digital archiving, where clearer images can enhance safety, usability, and analysis. However, as with many image enhancement techniques, there are potential risks if such technologies are misused to manipulate images, obscure evidence, or produce deceptive visual content. Additionally, improvements in image clarity might inadvertently reveal personal information or sensitive details in images that were previously obscured by poor quality, raising privacy concerns. We encourage future research to consider these ethical implications and to develop safeguards or detection mechanisms to identify manipulated or restored images. Our experiments focus on publicly available datasets, and we do not anticipate direct privacy risks from our work.

References

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1692–1700, 2018.

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

Abdullah Abuolaim, Mauricio Delbracio, Damien Kelly, Michael S Brown, and Peyman Milanfar. Learning to reduce defocus blur by realistically modeling dual-pixel data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2289–2298, 2021.

Mark R Banham and Aggelos K Katsaggelos. Digital image restoration. *IEEE signal processing magazine*, 14(2):24–41, 1997.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pp. 17–33. Springer, 2022.

Zuojun Chen, Pinle Qin, Jianchao Zeng, Quanzhen Song, Pengcheng Zhao, and Rui Chai. Lgit: local–global interaction transformer for low-light image denoising. *Scientific Reports*, 14(1):21760, 2024.

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/CVF international conference on computer vision*, pp. 4641–4650, 2021.

Xiaojie Chu, Liangyu Chen, Chengpeng Chen, and Xin Lu. Improving image restoration by revisiting global information aggregation. In *European Conference on Computer Vision*, pp. 53–71. Springer, 2022.

Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007.

Rui Deng and Tianpei Gu. Cu-mamba: Selective state space models with channel learning for image restoration. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 328–334. IEEE, 2024.

Feng Ding, Keping Yu, Zonghua Gu, Xiangjun Li, and Yunqing Shi. Perceptual enhancement for autonomous vehicles: Restoring visually degraded images for context prediction via adversarial training. *IEEE Transactions on Intelligent Transportation Systems*, 23(7):9430–9441, 2021.

Huiyu Duan, Wei Shen, Xiongkuo Min, Danyang Tu, Long Teng, Jia Wang, and Guangtao Zhai. Masked autoencoders as image processors. *arXiv preprint arXiv:2303.17316*, 2023.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision*, pp. 222–241. Springer, 2024.

Ali Hassani and Humphrey Shi. Dilated neighborhood attention transformer. *arXiv preprint arXiv:2209.15001*, 2022.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6185–6194, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Yindong Hua, Yifeng Liu, Binghan Li, and Mi Lu. Dilated fully convolutional neural network for depth estimation from a single image. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 612–616. IEEE, 2019.

Se-In Jang, Tinsu Pan, Ye Li, Pedram Heidari, Junyu Chen, Quanzheng Li, and Kuang Gong. Spach transformer: Spatial and channel-wise transformer based on local and global self-attentions for pet image denoising. *IEEE transactions on medical imaging*, 43(6):2036–2049, 2023.

Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8346–8355, 2020.

Ali Karaali and Claudio Rosito Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*, 27(3):1126–1137, 2017.

Junyong Lee, Sungkil Lee, Sunghyun Cho, and Seungyong Lee. Deep defocus map estimation using domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12222–12230, 2019.

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/CVF conference on computer vision and pattern recognition*, pp. 2034–2042, 2021.

Boyun Li, Haiyu Zhao, Wenxin Wang, Peng Hu, Yuanbiao Gou, and Xi Peng. Mair: A locality- and continuity-preserving mamba for image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition*, Nashville, TN, June 2025.

Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 254–269, 2018.

Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18278–18289, 2023.

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/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

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/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Chong Mou, Jian Zhang, and Zhuoyuan Wu. Dynamic attentive graph learning for image restoration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4328–4337, 2021.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3883–3891, 2017.

Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *European conference on computer vision*, pp. 327–343. Springer, 2020.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *CVPR*, 2022.

Kuldeep Purohit, Maitreya Suin, AN Rajagopalan, and Vishnu Naresh Boddeti. Spatially-adaptive image restoration using distortion-guided networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2309–2319, 2021.

Behnood Rasti, Yi Chang, Emanuele Dalsasso, Loic Denis, and Pedram Ghamisi. Image restoration for remote sensing: Overview and toolbox. *IEEE Geoscience and Remote Sensing Magazine*, 10(2):201–230, 2021.

Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3937–3946, 2019.

Ziyi Shen, Wenguan Wang, Xiankai Lu, Jianbing Shen, Haibin Ling, Tingfa Xu, and Ling Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5572–5581, 2019.

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*, pp. 657–665, 2015.

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

Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8174–8182, 2018.

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*, pp. 146–162. Springer, 2022.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11534–11542, 2020.

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/CVF conference on computer vision and pattern recognition*, pp. 17683–17693, 2022.

Wei Wei, Deyu Meng, Qian Zhao, Zongben Xu, and Ying Wu. Semi-supervised transfer learning for image rain removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3877–3886, 2019.

Rajeev Yasarla and Vishal M Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8405–8414, 2019.

Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*, 32, 2019.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pp. 492–511. Springer, 2020.

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/CVF conference on computer vision and pattern recognition*, pp. 14821–14831, 2021.

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/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Hai-Miao Zhang and Bin Dong. A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China*, 8(2):311–340, 2020.

He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 695–704, 2018.

Jiale Zhang, Yulun Zhang, Jinjin Gu, Yongbing Zhang, Linghe Kong, and Xin Yuan. Accurate image restoration with attention retractable transformer. *arXiv preprint arXiv:2210.01427*, 2022.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017a.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3929–3938, 2017b.

Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2737–2746, 2020.

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*, pp. 586–595, 2018.

Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, and Xi Peng. Comprehensive and delicate: An efficient transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14122–14132, 2023.