Research Article

Open Access

CrossMark

DiNAT-IR: Exploring dilated neighborhood attention for high-quality image restoration

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 introduce 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 upon acceptance.

Keywords: low-level computer vision, image restoration, transformer, neural networks

Volume 11 Issue 2 - 2025

Hanzhou Liu, Binghan Li, Chengkai Liu, Mi Lu
Electrical and Computer Engineering, Texas A&M University, USA

Correspondence: Mi Lu, Electrical and Computer Engineering, Texas A&M University, USA, Tel 979 845 3749

Received: August 23, 2025 | **Published:** September 23, 2025

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.[1–3] It aims to recover high-quality images from degraded inputs by addressing challenges like blur, noise, and other visual artifacts.[4] In autonomous driving and broader intelligent transportation systems (ITS), restoration techniques are crucial for enhancing visual inputs under adverse conditions, thereby supporting more reliable perception and decision-making.

In recent years, Transformers have emerged as powerful models for image restoration. Unlike traditional convolutional neural networks (CNNs) that rely on staked convolutional layers,[5–7] Transformers utilize self-attention to model long-range pixel relationships,[8–10] 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[10] 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.[11,12]

To bridge this gap, we explore Dilated Neighborhood Attention (DiNA) as a promising alternative, inspired by its recent success in detection and segmentation.[13] 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 fac-tors, effectively expanding the receptive field without incurring excessive computational overhead. The original DiNAT[14] reports that a hybrid design, using local neighborhood attention (NA) with a dilation factor δ = 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.

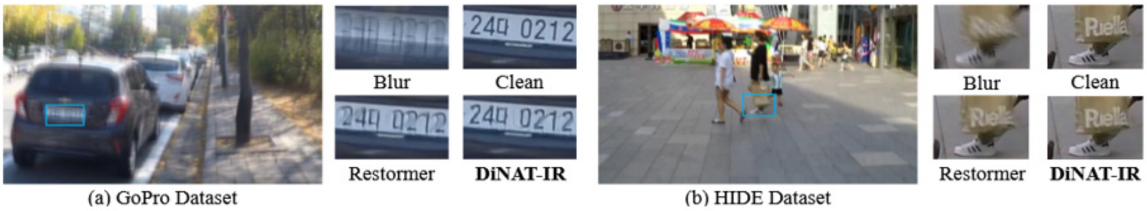


Figure 1 Visual comparisons between Restormer[10] and our proposed DiNAT-IR on the motion deblurring datasets.[15,16] DiNAT-IR produces cleaner restoration of numbers and characters on car license plates and hand-held bags. Zoom in to see details.

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 bottle-neck, 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:

a. We investigate the application of dilated neighborhood attention for image deblurring and identify key limitations of its hybrid attention design in this context.

b. We introduce a simple while effective channel-aware module that complements local neighborhood attention and restores global context without sacrificing pixel-level detail.

c. We propose DiNAT-IR, a Transformer-based architecture that achieves competitive performance not only on de-blurring benchmarks but also on other restoration tasks.

Related work

CNNs for Image Restoration. Convolutional neural networks (CNNs) have demonstrated strong performance across low-level computer vision tasks. DnCNN[5] pioneers the use of residual learning for image denoising, laying the groundwork for deeper and more effective architectures. MPRNet[6] adopts a multi-stage framework that processes image features at multiple spatial scales, achieving state-of-the-art (SOTA) results in image restoration. In the era of Transformer-based models, NAFNet[7] 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, CNN-based approaches typically rely on deeply stacked convolutional layers to enlarge the receptive field, which may restrict their ability to model long-range dependencies effectively (Figure 1).

Transformers for Image Restoration. Different from CNNs, Transformer-based architectures inherently model global context through self-attention mechanisms. While applying vanilla Transformers[17] 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[8] 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[9] 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[10] improves computational efficiency by computing self-attention along the channel dimension rather than spatial dimensions. However, follow-up studies[11,12] 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[18] and the Neighborhood Attention Transformer[19] 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)[13] 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.

Method

Preliminaries. In Transformer-based architectures, self-attention (SA) is the primary component responsible for most of the computational operations. Ours is an extension to Dilated Neighborhood Attention (DiNA).[13,19] DiNA serves as a flexible SA mechanism for short- and long-range learning by adjusting the dilation factor without additional complexity theoretically. For simplicity, consider a feature map X ∈ $R^{n \times d}$, where n is the number of tokens (basic data units) and d is the dimension; the query and key linear projections of X, Q and K; and relative positional biases between two tokens

i and j, B(i, j). Given a dilation factor δ and neighborhood size k, the attention weights for the i-th token is defined as:

i

$$A_i^{(k,\delta)} = \begin{bmatrix} Q_i K^T_{\rho_1(i)} + B_{(i,\rho_1^\delta(i))} \\ Q_i K^T_{\rho_2(i)} + B_{(i,\rho_2^\delta(i))} \\ \vdots \\ Q_i K^T_{\rho k(i)} + B_{(i,\rho_k^\delta(i))} \end{bmatrix} \quad (1)$$

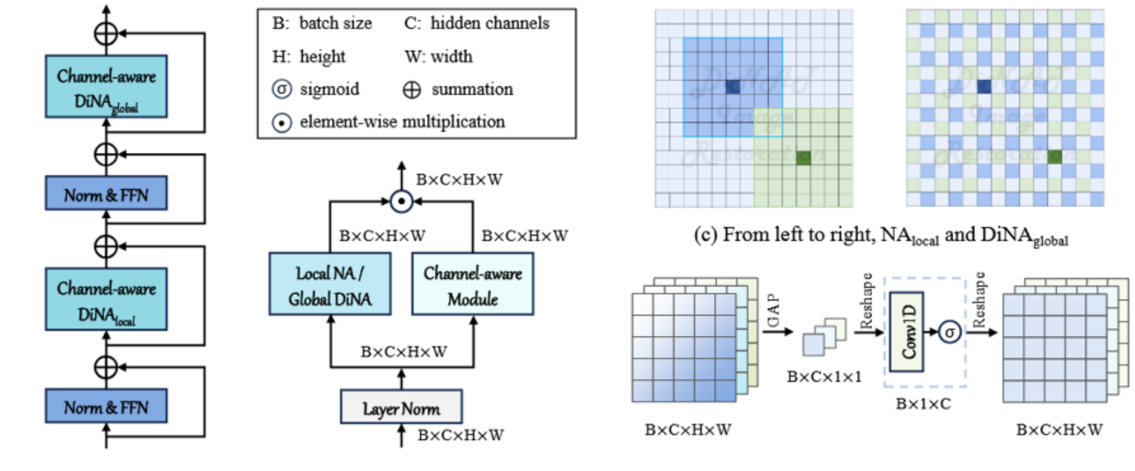
The corresponding matrix $V_i^{(k,\delta)}$, whose elements are the i-th token's k adjacent value linear projections, can be obtained by:

$$V_i^{(k,\delta)} = \begin{bmatrix} V^T_{\rho_1^\delta(i)} & V^T_{\rho_2^\delta(i)} & \cdots & V^T_{\rho_k^\delta(i)} \end{bmatrix}^T \quad (2)$$

The output feature map of DiNA for the i-th token is achieved by:

$$DiNA_k^\delta(i) = soft\max\left(\frac{A_i^{(k,\delta)}}{\sqrt{d_k}}\right) \mathrm{V}_i^{(k,\delta)} \quad (3)$$

Overall Pipeline. The overall pipeline of DiNAT-IR is based on Restormer.[10] 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 (Figure 2).



(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.

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 Trans-former (DiNAT).[13] By setting the dilation factor δ to 1, DiNA effectively reduces to standard Neighborhood Attention (NA).[19] 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 = 1 yields local NA, while larger values of δ expand the receptive field to capture broader context. The dilation pairs are defined as δ ∈ {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 computer vision tasks, such as motion deblurring, often 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.

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);[20,21] 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.

ˆ × ˆ × ˆ

Given a layer normalized input tensor X ∈ RH W C. The output of CASA is computed as,

$$\hat{X} = SA(X) \odot CAM(X),$$
$$CAM(X) = f^{-1}\left(Conv_{1d}\left(f\left(GAP_{2d}^{1\times1}(X)\right)\right)\right), \quad (4)$$

where ⊙ denotes element-wise multiplication; f is a tensor manipulation function which squeezes and transposes a C × 1 × 1 matrix, resulting in a 1 × C matrix; $Conv_{1d}$ denotes a 1D convolution with a kernel size of 3; $GAP^{1\times1}_{2d}$ indicates global average pooling, outputting a tensor of size 1 × 1. We employed the GAP design proposed by,[21] and the CAM idea draws inspiration from ECA-Net.[22]

Experiments and analysis

We evaluate the performance of DiNAT-IR across four distinct image restoration tasks: (a) single-image motion de-blurring, (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 for-matting 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[10] 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.[15] 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].[31] 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[10] and Stripformer.[32] 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[33] 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[34] 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 tsshe 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[3] 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.

Motion deblurring results

We conduct a thorough evaluation of various image restoration models on the GoPro[15] and HIDE[16] 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[23] and NAFNet.[7] 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[6] and attention-based method Restormer,[10] DiNAT-IR maintains similar or superior accuracy while pre-serving efficiency. Furthermore, despite being trained solely on GoPro, DiNAT-IR demonstrates excellent generalization to HIDE, underscoring 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 Figures 3&4 clearly demonstrate that the image deblurred by our method is more visually closer to the ground-truth than those of the other algorithms.

Table 1 Comparisons of image restoration models on gopro[15] and hide[16] datasets. we follow mair[23] 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[24]	30.26	0.934	28.36	0.904	3.76	35.87
DBGAN[25]	31.1	-	28.94	-	11.59	379.92
MT-RNN[26]	31.15	-	29.15	-	2.64	13.72
DMPHN[27]	31.2	-	29.09	-	86.8	-
CODE[28]	31.94	-	29.67	-	12.18	22.52
MIMO[29]	32.45	0.956	29.99	0.93	16.1	38.64
MPRNet[6]	32.66	0.958	30.96	0.939	20.13	194.42
Restormer[10]	32.92	0.961	31.22	0.942	26.13	35.31
Uformer[9]	33.06	0.967	30.9	0.953	50.88	22.36
CU-Mamba[30]	33.53	-	31.47	-	19.7	-
NAFNet[7]	33.69	0.966	31.32	0.942	67.89	15.85
MaIR[23]	33.69	0.969	31.57	0.946	26.29	49.29
DiNAT-IR	33.8	0.967	31.57	0.945	25.9	45.62

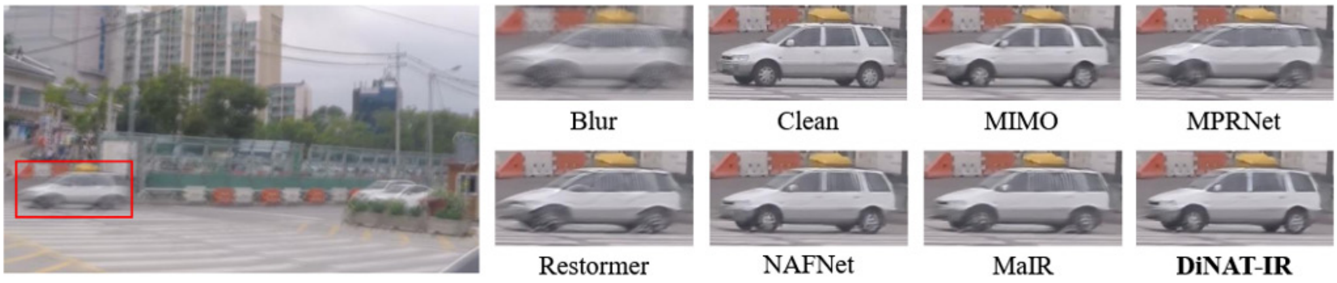


Figure 3 Single image motion deblurring results on the GoPro dataset [15]. Zoom in to see details.

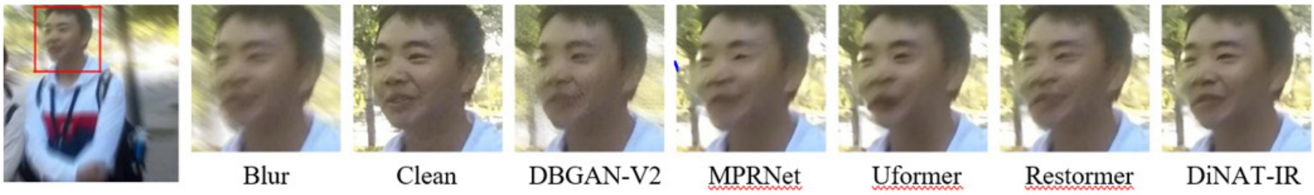


Figure 4 Single image motion deblurring results on the HIDE dataset [16]. Zoom in to see details.

Defocus deblurring results

Table 2 presents a comprehensive comparison of single-image and dual-pixel defocus deblurring methods on the DPDD dataset.[33] 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$_S$[36] and CSformer$_S$[42] 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.

Table 2 Dual-Pixel defocus deblurring comparisons on the DPDD dataset,[33] 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[36] across both single-image and dual-pixel settings.

Method	Indoor Scenes			Outdoor Scenes			Combined		
	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓	PSNR ↑	SSIM ↑	MAE ↓
EBDBS[37]	25.77	0.772	0.04	21.25	0.599	0.058	23.45	0.683	0.049
DMENetS[38]	25.5	0.788	0.038	21.43	0.644	0.063	23.41	0.714	0.051
JNBS[39]	26.73	0.828	0.031	21.1	0.608	0.064	23.84	0.715	0.048
DPDNetS[33]	26.54	0.816	0.031	22.25	0.682	0.056	24.34	0.747	0.044
KPACS[40]	27.97	0.852	0.026	22.62	0.701	0.053	25.22	0.774	0.04
IFANS[41]	28.11	0.861	0.026	22.76	0.72	0.052	25.37	0.789	0.039
RestormerS[10]	28.87	0.882	0.025	23.24	0.743	0.05	25.98	0.811	0.038
CSformerS[42]	29.01	0.883	0.023	23.63	0.759	0.047	26.25	0.819	0.036
GRL-BS[36]	29.06	0.886	0.024	23.45	0.761	0.049	26.18	0.822	0.037
DiNAT-IRS	28.94	0.881	0.025	23.48	0.751	0.049	26.14	0.814	0.037
DPDNetD[33]	27.48	0.849	0.029	22.9	0.726	0.052	25.13	0.786	0.041
RDPDD[43]	28.1	0.843	0.027	22.82	0.704	0.053	25.39	0.772	0.04
UformerD[9]	28.23	0.86	0.026	23.1	0.728	0.051	25.65	0.795	0.039
IFAND[41]	28.66	0.868	0.025	23.46	0.743	0.049	25.99	0.804	0.037
RestormerD[10]	29.48	0.895	0.023	23.97	0.773	0.047	26.66	0.833	0.035
CSformerD[42]	29.54	0.896	0.023	24.38	0.788	0.045	26.89	0.841	0.034
GRL-BD[36]	29.83	0.903	0.022	24.39	0.795	0.045	27.04	0.847	0.034
DiNAT-IRD	29.76	0.901	0.023	24.47	0.795	0.045	27.05	0.846	0.034

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.

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[10] 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,[44] DIDMDN,[45] and UMRL,[46] DiNAT-IR delivers significant improvements in both PSNR and SSIM. It also performs competitively against recent models like MPRNet[6] and SPAIR,[49] 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 3 Image deraining results. DINAT-IR achieves performance very close to that of restormer,[10] with ssim scores nearly matching those of restormer across multiple datasets. however, we acknowledge noticeably lower psnr scores on these datasets

	Rain100H		Rain100L		Test2800		Test1200		Test100	
Method	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
SEMI	16.56	0.486	25.03	0.842	24.43	0.782	26.05	0.822	22.35	0.788
DIDMDN	17.35	0.524	25.23	0.741	28.13	0.867	29.65	0.901	22.56	0.818
UMRL	26.01	0.832	29.18	0.923	29.97	0.905	30.55	0.91	24.41	0.829
RESCAN	26.36	0.786	29.8	0.881	31.29	0.904	30.51	0.882	25	0.835
PreNet	26.77	0.858	32.44	0.95	31.75	0.916	31.36	0.911	24.81	0.851
MSPFN	28.66	0.86	32.4	0.933	32.82	0.93	32.39	0.916	27.5	0.876
MPRNet	30.41	0.89	36.4	0.965	33.64	0.938	32.91	0.916	30.27	0.897
SPAIR	30.95	0.892	36.93	0.969	33.34	0.936	33.04	0.922	30.35	0.909
Restormer	31.46	0.904	38.99	0.978	34.18	0.944	33.19	0.926	32	0.923
DiNAT-IR	31.26	0.903	38.93	0.977	33.91	0.943	32.31	0.923	31.22	0.92

Denoising results

Table 4 compares several real-image denoising methods based on PSNR and SSIM metrics. Early approaches like DnCNN[50] [and BM3D[51] 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,[52] MIRNet,[53] MPRNet,[6] DAGL,[54] and Uformer[9] demonstrate significant improvements, achieving PSNR values around 39–40 dB and SSIM above 0.95, highlighting the advances brought by learning-based architectures. Among these methods, Restormer[10] achieves the highest performance, with a PSNR of 40.02 dB and an SSIM of 0.960, establishing a strong benchmark. MambaIR[55] 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.

Table 4 Real image denoising results. all methods are trained and tested on the sidd dataset,[35] Dinat-IR achieves performance comparable to restormer[10]

Method	**DnCNN[50]**	**BM3D[51]**	**VDN[52]**	**MIRNet[53]**	**MPRNet[6]**	**DAGL[54]**	**Uformer[9]**	**Restormer[10]**	**MambaIR[55]**	**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	0.96	0.96	0.96	0.96

Ablation study

In this section, we use Restormer[10] 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,[15] 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,[56,57] LPIPS[58] and NIQE, for comprehensive comparisons. Additionally, we report the total number of parameters and MACs to indicate complexity.

As shown in Table 5, the local-attention-only network already outperforms the Restormer baseline[10] 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[13] 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. 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.

Table 5 Ablation study on dilation factor configurations and the proposed channel-aware self-attention on the gopro[15] dataset. The baseline is restormer[10] 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	17.3
NA w/o CAM	31.56	0.948	11.12	0.111	3.0	16.64

Table 5 Continued....

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

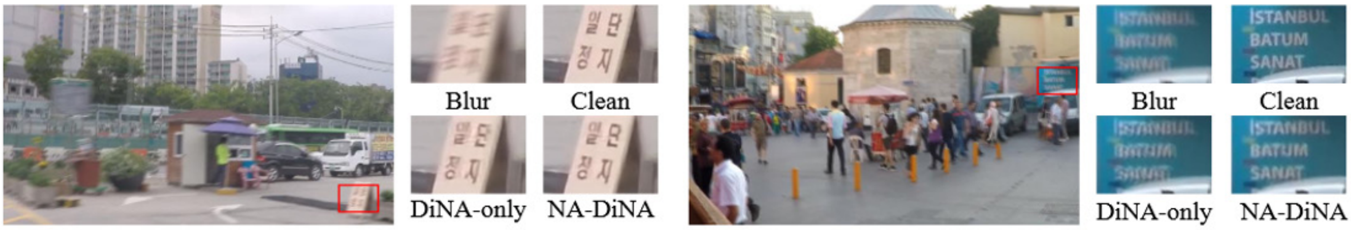


Figure 5 Visual comparisons between DiNAT-IR with global dilated neighborhood attention (DiNA) only and DiNAT-IR with both global DiNA and local neighborhood attention (NA). Both networks are trained and tested on the GoPro dataset[15] with the same training settings.

Limitation

While our approach demonstrates strong performance on standard benchmarks, there are several limitations worth noting. First, our ablation studies were primarily conducted on the GoPro dataset,[15] which ensures consistent training strategies across all models and allows for fair comparisons. How-ever, this dataset-specific focus may limit the generalizability of our findings to other restoration tasks and datasets. Extending the analysis to broader tasks is non-trivial, as it requires significant adaptation of training strategies to accommodate different data distributions and degradation characteristics.

Second, we observed training instability on certain datasets, such as those used for deraining tasks. This instability is potentially caused by dataset bias, which can hinder consistent model convergence. We believe that more robust and task-adaptive training strategies could alleviate this issue and further improve model performance. Overall, while our design presents a promising and effective direction for image restoration, a universally optimal architecture remains elusive due to the diverse and evolving nature of task requirements and dataset characteristics. Future work will focus on developing more specialized task-specific adaptations and designing generalized, robust training pipelines to further improve both the generalization ability and stability of image restoration models across a wide range of scenarios.

Conclusion

In this work, we investigated the challenges of applying Transformer-based attention to image restoration, highlighting the limitations of channel-wise attention and the difficulties of directly adopting dilated neighborhood attention (DiNA) for motion deblurring. To overcome these issues, we proposed DiNAT-IR, a Transformer framework that integrates DiNA with a lightweight channel-aware module. This design enhances global context modeling while preserving local detail, enabling more faithful recovery of fine textures and structures. Extensive experiments across multiple benchmarks confirm that DiNAT-IR achieves competitive results and consistently delivers perceptually clear restoration outputs.

Beyond deblurring, the architectural principles of DiNAT-IR demonstrate potential for broader image restoration tasks such as denoising and deraining. Given the importance of high-quality visual recovery in areas like autonomous driving, medical imaging, and remote sensing, we believe our frame-work provides both a practical solution and a foundation for future exploration. In particular, future work will investigate efficiency-oriented variants, and cross-domain generalization to unseen degradations.

Acknowledgements

None.

Conflicts of interest

Authors declares that there is no conflict of interest.

References

1. Ding F, Yu K, Gu Z, et al. Perceptual enhancement for autonomous vehicles: Restoring visually degraded images for context prediction via adversarial training. *IEEE Transactions on Intelligent Transportation Systems*. 2021;23(7):9430–9441, 2021.
2. Haimiao Zhang, Bin Dong. A review on deep learning in medical image reconstruction. *Journal of the Operations Research Society of China*. 2020;8(2):311–340.
3. Behnood Rasti; Yi Chang; Emanuele Dalsasso, et al. Image restoration for remote sensing: Overview and toolbox. IEEE Geoscience and Remote Sensing Magazine. 2021;10(2):201–230.
4. Banham MR, Katsaggelos AK. Digital image restoration. *IEEE signal processing magazine*. 1997;14(2):24–41.
5. Zhang K, Zuo W, Gu S, et al. Learning deep CNN denoiser prior for image restoration. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 3929–3938.
6. Zamir SW, Arora A, Khan S, et al. Multi-stage progressive image restoration. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, pp. 14 821–14 831.
7. Chen L, Chu X, Zhang X, et al. Simple baselines for image restoration. European conference on computer vision. Springer. 2022. p. 17–33.
8. Liang J, Cao J, Sun G, et al. Swinir: Image restoration using swin transformer. Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 1833–1844.
9. Wang Z, Cun X, Bao J, et al. Uformer: A general u-shaped transformer for image restoration. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, pp. 17 683–17 693.
10. Zamir SW, Arora A, Khan S, et al. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022. p. 5728–5739.
11. Jang SI, Pan T, Li Y, et al. Spach transformer: Spatial and channel-wise transformer based on local and global self-attentions for pet image denoising. *IEEE transactions on medical imaging*. 2023;43(6):2036–2049.

12. Chen Z, Qin P, Zeng J, et al. Lgit: local– global interaction transformer for low-light image denoising. *Scientific Reports*. 2024;14(1): 21760.

13. Hassani A, H Shi. Dilated neighborhood attention transformer.

14. Hua Y, Liu Y, Li B, et al. Dilated fully convolutional neural network for depth estimation from a single image. International Conference on Computational Science and Computational Intelligence (CSCI). IEEE. 2019. 612–616.

15. Nah S, Hyun Kim T, Mu Lee K. Deep multi-scale convolutional neural network for dynamic scene deblurring. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. 3883–3891.

16. Shen Z, Wang W, X Lu, et al. Human-aware motion deblurring. Proceedings of the IEEE/CVF international conference on computer vision, 2019. 5572–5581.

17. Vaswani. Attention is all you need. Advances in neural information processing systems. 2017.

18. Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision. 2021. 10 012–10 022.

19. Hassani S Walton, J Li, S Li, et al. Neighborhood attention transformer. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. 6185–6194.

20. Lin M, Q Chen, S Yan. Network in network.

21. Chu X, Chen L, Chen C, et al. Improving image restoration by revisiting global information aggregation," in European Conference on Computer Vision. Springer. 2022. 53–71.

22. Wang Q, Wu B, Zhu P, et al. Ecanet: Efficient channel attention for deep convolutional neural networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. 11 534–11 542.

23. B Li, H Zhao, W Wang, et al. "Mair: A locality- and continuity-preserving mamba for image restoration. IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, 2025.

24. Tao X, Gao H, Shen X, et al. Scale-recurrent network for deep image deblurring. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. 8174–8182.

25. Zhang K, Luo W, Zhong Y, et al. Deblurring by realistic blurring. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. 2737– 2746.

26. Park D, Kang DU, Kim J, et al. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training," in European conference on computer vision. Springer. 2020. 327–343.

27. Zhang J, Zhang Y, Gu J, et al. Accurate image restoration with attention retractable transformer.

28. Zhao H, Gou Y, Li B, et al. Comprehensive and delicate: An efficient transformer for image restoration. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

29. Cho SJ, Ji SW, Hong JP, et al. Rethinking coarse-to-fine approach in single image deblurring. In Proceedings of the IEEE/CVF international conference on computer vision. 2021. 4641–4650.

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

31. Loshchilov I, Hutter F. Decoupled weight decay regularization. 2017.

32. Tsai J, Peng YT, Lin YY, et al. Stripformer: Strip transformer for fast image deblurring. European conference on computer vision. Springer, 2022. p. 146–162.

33. Abuolaim A, Brown MS. Defocus deblurring using dual-pixel data. European Conference on Computer Vision. 2020. p. 111–126.

34. Jiang Z Wang, P Yi, C Chen, et al. Multi-scale progressive fusion network for single image deraining. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. p. 8346–8355.

35. Abdelhamed A, Lin S, Brown MS. A high-quality denoising dataset for smartphone cameras. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 1692–1700.

36. Y Li, Y Fan, X Xiang, et al. Efficient and explicit modelling of image hierarchies for image restoration. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023. p. 18278–18289.

37. Karaali A, CR Jung. Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing*. 2017;27(3):1126–1137.

38. Lee S, Lee S, Cho S, et al. Deep defocus map estimation using domain adaptation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019;12:222–12 230.

39. Shi J, Xu L, Jia J. Just noticeable defocus blur detection and estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015. p. 657–665.

1.

40. Son H, Lee J, Cho S, et al. Single image defocus deblurring using kernel-sharing parallel atrous convolutions. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021. p. 2642–2650.

41. Lee J, Son H, Rim J, et al. Iterative filter adaptive network for single image defocus deblurring. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021. p. 2034–2042.

42. Duan H, Shen W, Min X, et al. Masked autoencoders as image processors. 2023.

43. Abuolaim M, Delbracio D, Kelly MS, et al. Learning to reduce defocus blur by realistically modeling dual-pixel data. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. p. 2289–2298.

44. Wei W, Meng D, Zhao Q, et al. Semi-supervised transfer learning for image rain removal. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 3877– 3886.

45. Zhang H, VM 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. 2018. p. 695–704.

46. Yasarla R, VM 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. 2019. p. 8405–8414.

47. Li X, Wu J, Lin Z, et al. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In Proceedings of the European conference on computer vision (ECCV). 2018. p. 254–269.

48. Ren D, Zuo W, Hu Q, et al. Progressive image deraining networks: A better and simpler baseline. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019. p. 3937–3946.

49. Purohit K, Suin M, Rajagopalan A, et al. Spatially-adaptive image restoration using distortion-guided networks. In Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 2309–2319.

50. Zhang K, Zuo W, Chen Y, et al. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE transactions on image processing*. 2017;26(7):3142–3155.

51. Dabov K, Foi A, Katkovnik V, et al. Image denoising by sparse 3-d transform-domain collaborative filtering,. IEEE Transactions on image processing. 2007;16(8):2080–2095.

52. Yue Z, Yong H, Zhao Q, et al. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*. 2019;32.

53. Zamir SW, Arora A, Khan S, et al. 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. Springer; 2020. p. 492–511.

54. Mou C, Zhang J, Wu Z. Dynamic attentive graph learning for image restoration. In Proceedings of the IEEE/CVF international conference on computer vision. 2021. p. 4328–4337.

55. Guo H, Li J, Dai T, et al. Mambair: A simple baseline for image restoration with state-space model. European Conference on Computer Vision. Springer; 2024. p. 222– 241.

56. Heusel M, Ramsauer J, Unterthiner T, et al. Gans trained by a two time--scale update rule converge to a local NASH equilibrium," Advances in neural information processing systems. 2017;30.

57. Parmar G, Zhang R, JY. Zhu. On aliased resizing and surprising subtleties in GAN evaluation. CVPR, 2022.

58. Zhang R, Isola P, AA. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 586–595.