Fast and Accurate Image Restoration and Generation with Rank Enhanced Linear Attention

Yuang Ai

University of Chinese Academy of Sciences, Beijing, China
shallowdream555@gmail.com

Abstract. Transformer-based models have made remarkable progress in image restoration (IR) tasks. However, the quadratic complexity of self-attention in Transformer hinders its applicability to high-resolution images. Existing methods mitigate this issue with sparse or window-based attention, yet inherently limit global context modeling. Linear attention, a variant of softmax attention, demonstrates promise in global context modeling while maintaining linear complexity, offering a potential solution to the above challenge. Despite its efficiency benefits, vanilla linear attention suffers from a significant performance drop in IR, largely due to the low-rank nature of its attention map. To counter this, we propose Rank Enhanced Linear Attention (RELA), a simple yet effective method that enriches feature representations by integrating a lightweight depthwise convolution. Building upon RELA, we propose an efficient and effective Vision Transformer, named LAformer. LAformer eliminates hardware-inefficient operations such as softmax and window shifting, enabling efficient processing of high-resolution images. Extensive experiments across 7 IR tasks and 21 benchmarks demonstrate that LAformer outperforms SOTA methods and offers significant computational advantages. Furthermore, we extend LAformer to diffusion-based and flow-based visual generation, showcasing its strong potential as a competitive alternative to DiT and SiT. Code and models are available at https://github.com/shallowdream204/LAformer.

Keywords: Image Restoration · Linear Attention · Visual Generation

1 Introduction

Image restoration (IR) aims to recover a high-quality (HQ) image by removing degradations from its degraded low-quality (LQ) counterpart. As a fundamental task in low-level vision, image restoration has a wide array of real-world applications, spanning fields such as autonomous driving [129], surveillance monitoring [79], medical imaging [38], *etc*. Consequently, developing hardware-efficient IR networks holds substantial practical value.

Over the past decade, convolutional neural networks (CNNs) have achieved remarkable progress in various image restoration tasks [32, 106, 118]. However, CNNs are often limited by their constrained receptive field, making it challenging

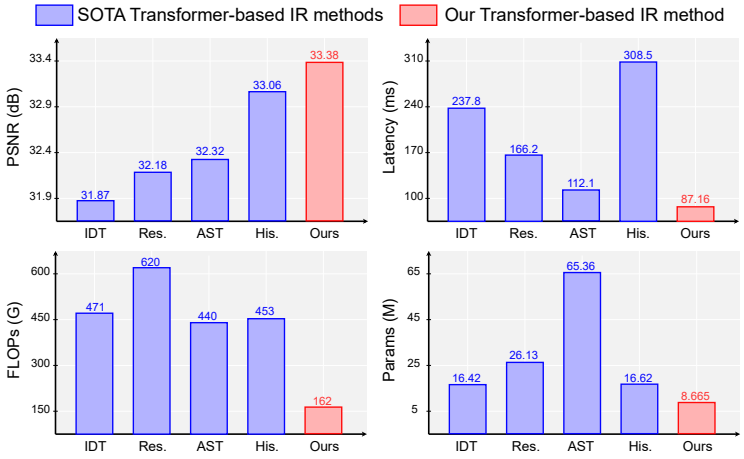

Fig. 1: Comprehensive comparison with SOTA Transformer-based methods for image restoration (raindrop removal).

to capture long-range dependencies. Recently, Transformer-based methods [9,24, 113] have significantly advanced the performance of image restoration.

While Transformer offers the advantage of a global receptive field, its core self-attention mechanism introduces quadratic computational complexity. Directly applying self-attention globally incurs substantial overhead, particularly when handling high-resolution images. To mitigate this, some methods compute self-attention within local windows [23, 53, 99], while others adopt sparse attention patterns [22,116]. While effectively reducing computational costs, these methods either limit global perception or are sensitive to degradation types due to specific attention patterns [126].

The above analysis leads to a crucial question: *Is it possible to design a hardware-efficient image restoration model that simultaneously preserves strong global perceptive capabilities?* In this paper, we explore linear attention [42] as a potential solution to this issue. Compared to standard softmax attention [91], linear attention replaces the softmax operation with simpler activation functions (*e.g.*, ReLU [3], ELU [27]) and reorders the computation to first calculate $K^\top V$. As shown in Fig. 2, linear attention reduces the computational complexity to a linear scale with respect to the number of visual tokens N, while preserving the global contextual advantages of softmax attention. When processing high-resolution images, the global receptive field of linear attention enables the capture of long-range relationships between pixels, thereby enhancing restoration quality. These attributes motivate us to employ linear attention in developing an efficient and effective image restoration model.

However, standard linear attention mechanisms appear suboptimal for image restoration, resulting in a notable performance drop relative to softmax attention (see Tab. 14). Through rank analysis [52], we find that the low-rank nature of the attention map in linear attention severely limits the diversity of its out-

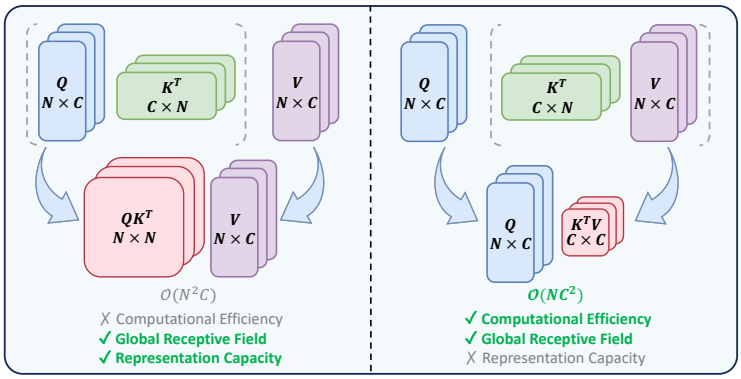

Fig. 2: Conceptual comparison between **Softmax Attention** and **Linear Attention**. While linear attention excels at efficiently capturing global context, its representation capacity is limited, resulting in a notable performance drop.

put features, thereby diminishing its representation capacity (see Fig. 5). To overcome this limitation, we propose Rank Enhanced Linear Attention (RELA), which incorporates a lightweight depthwise convolution to effectively enhance the rank of output feature. RELA retains linear complexity while boosting the representation capacity of linear attention, enabling improved modeling of complex degradation patterns.

We further propose LAformer, an efficient Vision Transformer structured within the widely adopted U-Net architecture [104, 113]. As shown in Fig. 3, LAformer incorporates a Dual-Attention (DA) Block that operates across spatial and channel dimensions, leveraging both RELA and channel attention [121] concurrently to facilitate efficient global information capture. Recognizing the crucial role of local information in restoring texture details, we introduce a Convolutional Gated Feed-Forward Network (CG-FFN) to strengthen LAformer's capability in modeling spatial locality. This architecture enables LAformer to effectively integrate local and global information, enhancing its representational power. Notably, LAformer circumvents hardware-inefficient operations like softmax and window shifting, significantly boosting its efficiency for high-resolution image processing. As shown in Fig. 1, our LAformer surpasses SOTA Transformer-based IR methods and significantly reduces computational overhead. Furthermore, we extend LAformer for diffusion and flow-based visual generation, demonstrating its remarkable efficiency and scalability.

Our main contributions can be summarized as follows:

- We propose LAformer, a simple, efficient, and strong Vision Transformer that combines powerful representation capability with linear complexity.
- We propose Rank Enhanced Linear Attention (RELA), a simple yet effective mechanism that boosts feature diversity in linear attention by enhancing the rank of features.

- Extensive experiments across 7 IR tasks and 21 benchmarks validate that LAformer achieves SOTA performance while maintaining high efficiency.
- When extended to visual generation, LAformer demonstrates impressive efficiency and scalability, highlighting its potential as a competitive alternative to DiT and SiT.

2 Related Work

Image Restoration. Images captured in real-world scenarios often suffer from various degradations, such as blur, low-light, and adverse weather conditions. These issues inevitably degrade the performance of downstream models and diminish the effectiveness of many vision systems [4, 10, 65, 129]. SRCNN [32] and DnCNN [118] are pioneering works that introduce CNN into image restoration and achieve impressive results. Subsequently, a plethora of CNN-based architectures incorporating meticulously designed modules have been introduced to enhance IR performance. Notable examples include the residual block [43,47,54], dense block [97,122,123], channel attention [121], and non-local attention mechanisms [67, 68, 103]. Despite their success, CNNs remain constrained by their inherently limited receptive field, posing challenges in effectively capturing long-range dependencies. Recently, researchers have begun to use self-attention mechanisms in place of convolutions to enhance IR performance [18, 53, 99, 113].

Vision Transformer. Since its introduction by NLP researchers, Transformer [91] has been widely adopted in computer vision tasks [6,34,36,37,60]. IPT [18] is the pioneering work that introduces Transformer to image restoration by processing images in small patches. Since then, many works [16, 52, 113, 119, 124] have focused on improving self-attention, aiming to reduce its significant computational overhead. SwinIR [53] and Uformer [99] perform self-attention in non-overlapping local windows. HAT [23] proposes overlapping cross-attention to establish cross-window interaction. ART [116] introduces sparse attention to IR for a large receptive field. AST [126] proposes an adaptive sparse self-attention to filter out irrelevant tokens. Although effective, a trade-off between global receptive field and computational efficiency persists.

Due to its efficiency and ability to capture global context, Linear attention [42] has garnered significant interest from the computer vision community. Shen *et al*. [82] introduce linear attention to object detection and instance segmentation. Bolya *et al*. [13] propose hydra attention by maximizing the number of linear attention heads. Han *et al*. [41] propose a focused linear attention module to enhance efficiency and expressiveness. EfficientViT [14] finds that linear attention cannot produce sharp attention distributions and proposes multi-scale linear attention to enhance local information extraction. Distinct from prior works, ours is the first to demonstrate the efficacy of linear attention in high-resolution image restoration. Additionally, we investigate the limitation of linear attention in representation capacity and propose novel approaches to overcome the challenge.

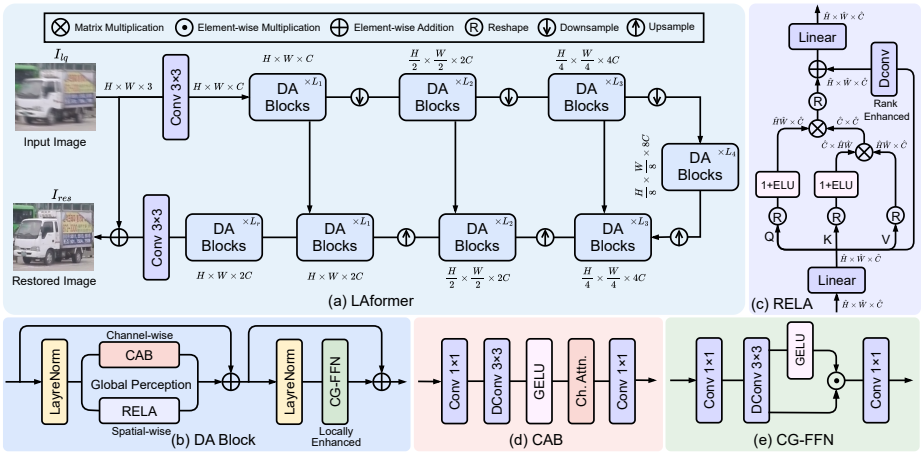

Fig. 3: Overall architecture of LAformer, which includes (b) Dual Attention (DA) Block, (c) Rank Enhanced Linear Attention (RELA), (d) Channel Attention Block (CAB), and (e) Convolutional-Gated Feed-Forward Network (CG-FFN).

3 Method

We propose LAformer, an efficient framework for high-resolution IR. First, we analyze the softmax attention and linear attention mechanisms in Sec. 3.1. Then Sec. 3.2 delves into the inherent low-rank characteristics of linear attention, motivating the development of Rank Enhanced Linear Attention (RELA) as a solution. Finally, in Sec. 3.3, we present LAformer, a model built on RELA that achieves explicit global modeling with linear complexity.

3.1 Softmax vs. Linear Attention

Softmax attention [91] is a widely used mechanism in Transformer. For a given input $X \in \mathbb{R}^{N\times C}$, where N represents the number of visual tokens and C denotes the channel dimension, it can be generally formulated as

$$Y_i = \sum_{j=1}^{N} \frac{\mathrm{Sim}(Q_i, K_j)}{\sum_{j=1}^{N} \mathrm{Sim}(Q_i, K_j)} V_j, \tag{1}$$

where $Q = XW_Q, K = XW_K, V = XW_V$, $W_{Q,K,V} \in \mathbb{R}^{C\times C}$ are learnable linear projection matrices, $\mathrm{Sim}(\cdot,\cdot)$ is the similarity function, i, j index visual tokens. When the similarity function is set as $\mathrm{Sim}(Q_i, K_j) = \exp(\frac{Q_i K_j^\top}{\sqrt{C}})$, Eq. (1) represents the standard softmax attention. Due to the need to compute similarities between all queries and keys, softmax attention incurs a $\mathcal{O}(N^2)$ complexity.

Linear attention [42] uses simple activation functions to approximate the similarity function as $\mathrm{Sim}(Q_i, K_j) = \psi(Q_i)\psi(K_j)^\top$, where $\psi(\cdot)$ is the activation

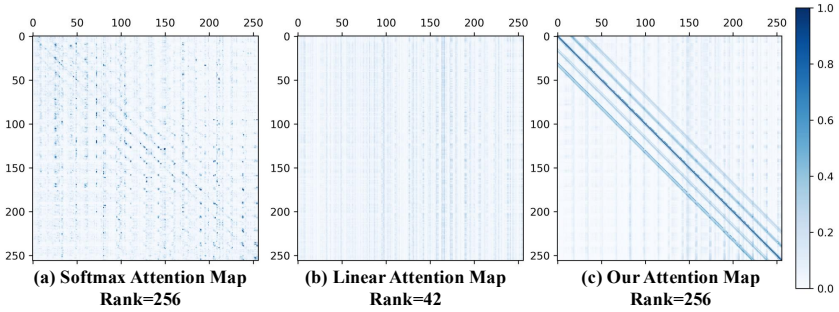

Fig. 4: Visualization of attention maps ($N = 256, C = 48$) of softmax attention (window-based), linear attention, and RELA. Our RELA effectively restores the rank of the attention map.

function. With this assumption, we can rewrite Eq. (1) as

$$\begin{aligned} Y_i &= \sum_{j=1}^{N} \frac{\psi(Q_i)\psi(K_j)^\top}{\sum_{j=1}^{N} \psi(Q_i)\psi(K_j)^\top} V_j \\ &= \frac{\psi(Q_i)\left(\sum_{j=1}^{N} \psi(K_j)^\top V_j\right)}{\psi(Q_i)\left(\sum_{j=1}^{N} \psi(K_j)^\top\right)}. \end{aligned} \tag{2}$$

Compared to softmax attention, linear attention modifies the computation order from $(QK^\top)V$ to $Q(K^\top V)$, reducing the complexity from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ while preserving the global modeling capabilities. However, the improvement in computational efficiency is accompanied by a noticeable drop in performance (see Tab. 14).

3.2 Rank Enhanced Linear Attention

Following [52], we conduct rank analysis for the linear attention map M, which is computed as $\psi(Q)\psi(K)^\top \in \mathbb{R}^{N\times N}$. According to basic matrix theory, it follows that

$$\begin{aligned} \text{Rank}(M) &\leq \min(\text{Rank}(\psi(Q)), \text{Rank}(\psi(K)^\top) \\ &\leq \min(N, C). \end{aligned} \tag{3}$$

For high-resolution IR, the number of visual tokens N is typically much larger than C, which results in the low-rank nature of the attention map in linear attention (see Fig. 4). Since the attention output is a weighted sum of V using the attention map, the highly correlated weights inevitably limit the expressive capacity of the output features. As shown in Fig. 5, the rank of output features produced by linear attention is notably lower than that of softmax attention, suggesting a reduced diversity in features.

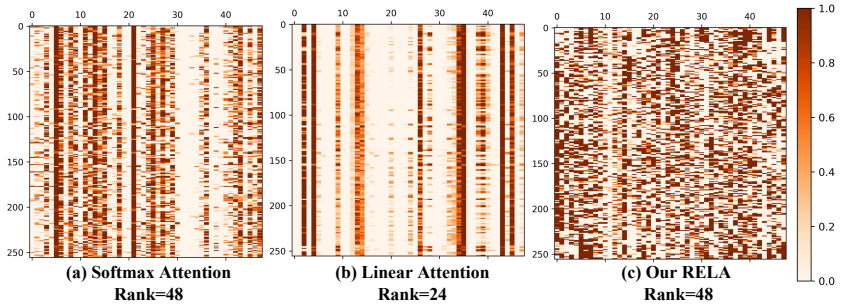

Fig. 5: Visualization of feature maps output by softmax attention (window-based), linear attention, and RELA. Our RELA substantially enriches the expressiveness of the resulting features.

To address this limitation, we introduce Rank Enhanced Linear Attention (RELA), depicted in Fig. 3 (c). We integrate a lightweight depth-wise convolution (DWC) into the computation process, which can be formulated as

$$Y = (1 + \mathrm{ELU}(Q))(1 + \mathrm{ELU}(K))^\top V + W_d V, \tag{4}$$

where W_d represents the depth-wise convolution. Following [42], we use $1 + \mathrm{ELU}(\cdot)$ as the activation function.

The DWC can be interpreted as a local attention operation, enriching the output through local feature combinations. We can rewrite Eq. (4) as $Y = (\psi(Q)\psi(K)^\top + M_{DWC})V = M_{attn}V$, where M_{DWC} represents the sparse matrix derived from the DWC operation, and M_{attn} denotes the resulting full attention matrix. Since M_{DWC} can potentially be of full rank, its incorporation effectively raises the upper bound of the rank of the overall attention map M_{attn}. Fig. 4 and Fig. 5 illustrate that our RELA successfully restores the attention map and output features to a full-rank state, substantially enhancing the diversity of learned features. It is worth noting that our RELA introduces only a minimal increase in parameters and computational overhead compared to standard linear attention, yet it brings a significant performance boost (see Tab. 14).

3.3 Network Architecture

Overall Architecture The overall architecture of our LAformer is depicted in Fig. 3. To rigorously assess the effectiveness of our proposed method, we adopt the classical U-Net architecture as the backbone, aligning with previous approaches [19, 99, 113]. For an input low-quality (LQ) image $I_{lq} \in \mathbb{R}^{H\times W\times 3}$, LAformer initially applies a 3×3 convolution for overlapping patch embedding, transforming I_{lq} into a feature map with C channels. In both the encoder and decoder modules, we incorporate Dual Attention (DA) Blocks to efficiently capture both global and local degradation patterns. Within each stage, skip connections

link the encoder and decoder, enabling seamless information flow across intermediate features. Across stages, pixel-unshuffle and pixel-shuffle operations are employed for effective feature down-sampling and up-sampling.

Dual Attention Block To improve both global and local feature extraction in LAformer, we integrate a dual attention module and a Convolutional Gated Feed-Forward Network (CG-FFN), which address global and local feature learning, respectively. The details of these components are outlined below.

Dual Attention. We use the proposed RELA to achieve efficient global modeling in the spatial dimension. Considering that channel-wise information also plays an important role in IR, we introduce channel attention [121] into the DA Block, forming the Channel Attention Block (CAB) in conjunction with convolutions, as depicted in Fig. 3 (d). The process of CAB can be formulated as

$$Y = W_{p_2}\mathrm{CA}(\mathrm{GELU}(W_d W_{p_1} X)), \tag{5}$$

where $W_{p_{(\cdot)}}$ is the 1×1 point-wise convolution, W_d is the 3×3 depth-wise convolution, $\mathrm{CA}(\cdot)$ denotes the channel attention operation. Channel attention can be interpreted as a mechanism for modeling global dependencies, as it aggregates token information across the spatial domain through global average pooling. In the DA Block, we combine RELA and CAB to effectively model and aggregate global information across both spatial and channel dimensions, achieving efficient global modeling and aggregation.

Convolutional Gated FFN. As discussed in [14], compared to softmax attention, linear attention tends to produce less sharp attention maps, which hinders its ability to effectively capture local information. However, rich local information is crucial for restoring texture details. To address this, we propose a simple Convolutional Gated Feed-Forward Network (CG-FFN) to enhance the local fitting capability of LAformer. As illustrated in Fig. 3 (e), CG-FFN integrates depth-wise convolution with Gated Linear Unit (GLU) [30], which can be formulated as

$$Y = W_{p_2}(\mathrm{GELU}(W_d W_{p_1} X) \odot W_d W_{p_1} X). \tag{6}$$

Compared to MLP, CG-FFN incorporates depth-wise convolution to enhance LAformer's capacity for capturing fine-grained local details while reducing computational cost, thereby complementing the linear attention mechanism.

Dual-Attention Block. Our DA Block is equipped with RELA, CAB, and CG-FFN, which is defined as

$$\begin{aligned} X' &= \mathrm{RELA}(\mathrm{LN}(X)) + \mathrm{CAB}(\mathrm{LN}(X)) + X, \\ Y &= \mathrm{CGFFN}(\mathrm{LN}(X')) + X'. \end{aligned} \tag{7}$$

Built upon the stacked DA Blocks, LAformer avoids hardware-inefficient operations (*e.g.*, softmax, window shifting) and explicitly models spatial-wise global features, channel-wise global features, and local features, all with linear complexity. This enables LAformer to maintain high efficiency while possessing strong representation capability.

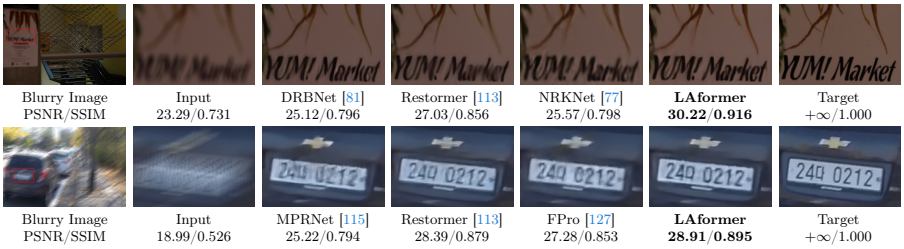

Fig. 6: Top row: Visual comparison on the DPDD [1] dataset for **single-image defocus deblurring**. Bottom row: Visual comparison on the GoPro [69] dataset for **motion deblurring**. Zoom in for a better view.

4 Experiments

4.1 Experimental Setup

We evaluate the proposed LAformer on various datasets for **7 image restoration tasks**: **(1)** motion deblurring, **(2)** defocus deblurring, **(3)** image dehazing, **(4)** image desnowing, **(5)** raindrop removal, **(6)** low-light enhancement, and **(7)** all-in-one image restoration. Furthermore, we extend LAformer for **diffusion-based and flow-based visual generation**. Additional details about the training dataset, training protocols, and more visual results can be found in the **Appendix** due to space limits.

Implementation Details. The kernel size for the depth-wise convolution in RELA is set to 5×5. To align with SOTA models for each specific task, we adapt the number of blocks and channels, considering variants of different sizes, including Tiny, Small, and Base versions (LAformer-T/S/B). All experiments are conducted on A100 GPUs. Additional task-specific training hyperparameters and details of the model architecture can be found in the **Appendix**.

4.2 Image Restoration

Motion Deblurring. Tab. 2 and Tab. 1 show the motion deblurring results on the synthetic and real-world datasets, respectively. When trained on the synthetic GoPro [69] dataset, our model outperforms all comparison methods on both the GoPro and HIDE [83] datasets. When trained on the real RealBlur [80] dataset, our model achieves the best results, surpassing GRL [52] by 0.87 dB in

Table 1: Motion deblurring results on real-world datasets. All models are trained and tested on RealBlur [80].

Dataset	Method	DeblurGAN-v2 [46]	SRN [85]	MIMO-UNet+ [26]	MPRNet [115]	DeepRFT+ [88]	BANet [44]	MSSNet [66]	Stripformer [87]	FFTformer [45]	GRL [52]	LAformer-B Ours
RealBlur-R [80]	PSNR	36.44	38.65	N/A	39.31	39.55	39.76	39.84	39.84	40.11	40.20	**41.07**
	SSIM	0.935	0.965	N/A	0.972	0.971	0.972	0.972	0.974	0.973	0.974	**0.977**
RealBlur-J [80]	PSNR	29.69	31.38	31.92	31.76	32.00	32.10	32.19	32.48	32.62	32.82	**32.92**
	SSIM	0.870	0.909	0.919	0.922	0.923	0.928	0.931	0.929	0.933	0.932	**0.933**

Table 2: Motion deblurring results on synthetic datasets.

Method	Venue	GoPro [69]		HIDE [83]	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
MIMO-Unet+ [26]	ICCV'21	32.45	0.957	29.99	0.930
IPT [18]	CVPR'21	32.52	N/A	N/A	N/A
MPRNet [115]	CVPR'21	32.66	0.959	30.96	0.939
Restormer [92]	CVPR'22	32.92	0.961	31.22	0.942
Stripformer [87]	ECCV'22	33.08	0.962	31.03	0.940
HI-Diff [25]	NIPS'23	33.33	0.964	31.46	0.945
IRNeXt [29]	ICML'23	33.16	0.962	N/A	N/A
PromptRestorer [92]	NIPS'23	33.06	0.962	31.36	0.944
FPro [127]	ECCV'24	33.05	0.961	N/A	N/A
ACL [39]	CVPR'25	33.25	0.964	N/A	N/A
LAformer-B	-	**33.40**	**0.965**	**31.49**	**0.945**

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

Method	CSD [21]		SRRS [20]		Snow100K [58]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DesnowNet [58]	20.13	0.81	20.38	0.84	30.50	0.94
CycleGAN [35]	20.98	0.80	20.21	0.74	26.81	0.89
All in One [51]	26.31	0.87	24.98	0.88	26.07	0.88
JSTASR [20]	27.96	0.88	25.82	0.89	23.12	0.86
HDCW-Net [21]	29.06	0.91	27.78	0.92	31.54	0.95
TransWeather [90]	31.76	0.93	28.29	0.92	31.82	0.93
NAFNet [19]	33.13	0.96	29.72	0.94	32.41	0.95
FocalNet [28]	37.18	0.99	31.34	0.98	33.53	0.95
LAformer-T	**37.42**	**0.99**	**32.24**	**0.98**	**34.24**	**0.96**

Table 4: Model efficiency comparison with Transformer-based methods on the GoPro [69] dataset (1280 × 720 resolution).

Method	Uformer [99]	Restormer [113]	Stripformer [87]	HI-Diff [25]	**LAformer-B Ours**
#Params (M)	50.88	26.13	19.71	28.49	24.89
FLOPs (G)	86.89	154.88	155.03	142.62	144.33
Runtime (s)	1.41	0.79	0.87	1.56	0.73
PSNR (dB)	32.97	32.96	33.08	33.33	33.40

PSNR on RealBlur-R. Fig. 6 shows that LAformer produces sharper and clearer results compared to other methods.

We present a comparison of model efficiency on the GoPro dataset in Tab. 4, where FLOPs are computed using 256×256 input images. Our method demonstrates superior efficiency compared to other approaches.

Defocus Deblurring. Tab. 5 shows the comparison results on the DPDD [1] dataset. Our LAformer consistently outperforms SOTA methods across all three scene types. Specifically, in the combined scenes, LAformer achieves a PSNR improvement of 0.3 dB over the Transformer-based method Restormer [113] in the single-image setting, and a 0.29 dB improvement in the dual-pixel setting. Fig. 6 shows that LAformer restores clear text, whereas other methods result in noticeable blurring or artifacts.

Image Dehazing. Tab. 6 and Tab. 7 present the dehazing results on the synthetic and real-world datasets, respectively. LAformer achieves the best performance for all datasets. Fig. 7 shows that LAformer achieves the dehazing result closest to the ground truth compared to other methods.

Image Desnowing. Tab. 3 shows the desnowing results on three widely used datasets. We follow the dataset split used in FocalNet [28]. Our method achieves the best performance across three datasets.

Raindrop Removal. Tab. 8 shows the raindrop removal results on the Raindrop [73] dataset. Compared to recent methods AST [126] and Histoformer [84], LAformer achieves a PSNR gain of 1.06 dB and 0.32 dB, respectively.

Low-light Enhancement. As shown in Tab. 9, LAformer achieves the best PSNR score for all datasets. Fig. 7 shows that LAformer produces the clearest

Table 5: Defocus deblurring results on the DPDD testset [1]. **S:** single-image defocus deblurring. **D:** dual-pixel defocus deblurring.

Method	Indoor Scenes				Outdoor Scenes				Combined			
	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓	PSNR ↑	SSIM ↑	MAE ↓	LPIPS ↓
DRBNet$_S$ [81]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	25.73	0.791	N/A	0.183
Restormer$_S$ [113]	28.87	0.882	0.025	**0.145**	23.24	0.743	0.050	0.209	25.98	0.811	0.038	0.178
NRKNet$_S$ [77]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	26.11	0.810	N/A	0.210
INIKNet$_S$ [78]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	26.05	0.803	N/A	0.185
MPerceiver$_S$ [8]	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	25.88	0.803	N/A	0.190
LAformer-B$_S$	**29.13**	**0.886**	**0.024**	0.146	**23.57**	**0.759**	**0.048**	**0.207**	**26.28**	**0.821**	**0.036**	**0.177**
RDPD$_D$ [2]	28.10	0.843	0.027	0.210	22.82	0.704	0.053	0.298	25.39	0.772	0.040	0.255
Uformer$_D$ [99]	28.23	0.860	0.026	0.199	23.10	0.728	0.051	0.285	25.65	0.795	0.039	0.243
IFAN$_D$ [48]	28.66	0.868	0.025	0.172	23.46	0.743	0.049	0.240	25.99	0.804	0.037	0.207
Restormer$_D$ [113]	29.48	0.895	0.023	0.134	23.97	0.773	0.047	0.175	26.66	0.833	0.035	0.155
LAformer-B$_D$	**29.53**	**0.898**	**0.023**	**0.125**	**24.49**	**0.795**	**0.044**	**0.160**	**26.95**	**0.845**	**0.034**	**0.143**

Table 6: Image dehazing results on synthetic datasets.

Method	SOTS-Indoor [49]		SOTS-Outdoor [49]	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
GridDehazeNet [56]	32.16	0.984	30.86	0.982
MSBDN [33]	33.67	0.985	33.48	0.982
FFA-Net [74]	36.39	0.989	33.57	0.984
AECR-Net [101]	37.17	0.990	N/A	N/A
MAXIM [89]	38.11	0.991	34.19	0.985
DeHamer [40]	36.63	0.988	35.18	0.986
PMNet [112]	38.41	0.990	34.74	0.985
FocalNet [28]	40.82	0.996	37.71	0.995
MB-TaylorFormer [75]	40.71	0.992	37.42	0.989
LAformer-T	**41.17**	**0.996**	**38.51**	**0.995**

Table 7: Image dehazing results on real-world datasets.

Method	Outdoor-Haze [12]		Dense-Haze [11]	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
GridDehazeNet [56]	18.92	0.672	14.96	0.536
MSBDN [33]	24.36	0.749	15.13	0.555
FFA-Net [74]	22.12	0.770	15.70	0.549
AECR-Net [101]	N/A	N/A	15.80	0.470
Restormer [113]	23.58	0.768	15.78	0.548
DeHamer [40]	25.11	0.777	16.62	0.560
PMNet [112]	24.64	0.830	16.79	0.510
MB-TaylorFormer [75]	25.05	0.788	16.66	0.560
AST [126]	N/A	N/A	17.12	0.550
LAformer-T	**25.28**	**0.932**	**17.14**	**0.565**

visual results, while the results of the comparative methods contain noticeable noise and artifacts.

All-in-One IR. We further validate the effectiveness of LAformer on the all-in-one IR task. Following the five-task setting of [7,125], we train a single LAformer model to address five different types of degradations. As shown in Tab. 10, our LAformer outperforms DiffUIR across all tasks, delivering significant performance gains. This further highlights the strong representation capacity of LAformer, which achieves superior results without requiring any specialized design for the all-in-one IR task.

4.3 Visual Generation

Following [5,64,71], we conduct class-conditional image generation experiments on ImageNet [31] at 256×256 resolution. We adjust the depth and width of LAformer to match the parameter count and computational cost of baselines. ***We follow exactly the same experimental setup as DiT and SiT to ensure a fair comparison.*** Tab. 11 shows that across models of varying scales (ranging from approximately 33M to 680M parameters), our LAformer achieves superior generation performance with lower computational cost compared to DiT. This

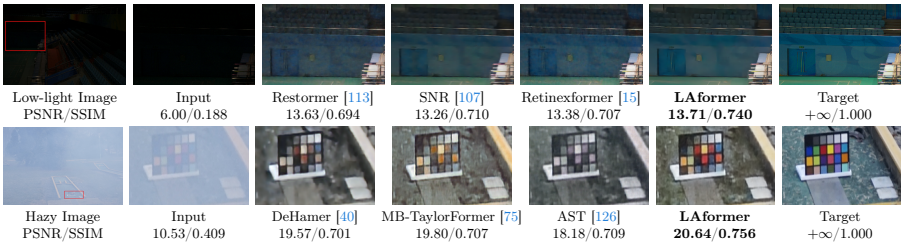

Fig. 7: Top row: Visual comparison on the LOL-v1 [100] dataset for **low-light enhancement**. Bottom row: Visual comparison on the Dense-Haze [11] dataset for **image dehazing**. Zoom in for a better view.

Table 8: Raindrop removal results on the Raindrop [73] dataset.

Method	Venue	Test-A		Test-B	
		PSNR	SSIM	PSNR	SSIM
AGAN [73]	CVPR'18	31.62	0.921	25.05	0.811
Quan *et al.* [76]	ICCV'19	31.44	0.926	N/A	N/A
DuRN [57]	CVPR'19	31.24	0.926	25.32	0.817
Restormer [113]	CVPR'22	32.18	0.941	N/A	N/A
MAXIM [89]	CVPR'22	31.87	0.935	25.74	0.827
IDT [105]	TPAMI'23	31.87	0.931	N/A	N/A
WeatherDiff [70]	TAPMI'23	32.43	0.933	N/A	N/A
AWRCP [110]	ICCV'23	31.93	0.931	N/A	N/A
DTPM [111]	CVPR'24	32.72	0.944	N/A	N/A
AST [126]	CVPR'24	32.32	0.935	N/A	N/A
Histoformer [84]	ECCV'24	33.06	0.944	N/A	N/A
MOERL [95]	ICCV'25	33.21	0.945	N/A	N/A
LAformer-S	-	**33.38**	**0.949**	**27.27**	**0.850**

Table 9: Low-light enhancement results on three datasets.

Method	LOL-v1 [100]		LOL-v2-real [109]		LOL-v2-syn [109]	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
SID [17]	14.35	0.436	13.24	0.442	15.04	0.610
IPT [18]	16.27	0.504	19.80	0.813	18.30	0.811
Uformer [99]	16.36	0.771	18.82	0.771	19.66	0.871
Sparse [109]	17.20	0.640	20.06	0.816	22.05	0.905
RUAS [55]	18.23	0.720	18.37	0.723	16.55	0.652
SCI [63]	14.78	0.646	20.28	0.752	24.14	0.928
KinD [120]	20.87	0.802	17.54	0.669	13.29	0.578
MIRNet [114]	24.14	0.830	20.02	0.820	21.94	0.876
DRBN [108]	19.86	0.834	20.13	0.830	23.22	0.827
URetinex-Net [102]	21.33	0.835	21.16	0.840	24.14	0.928
Restormer [113]	22.43	0.823	19.94	0.827	21.41	0.830
MRQ [59]	25.24	0.855	22.37	0.846	25.94	0.935
LLFlow [98]	25.19	**0.870**	26.53	**0.892**	26.23	0.943
SNR [107]	24.61	0.842	23.92	0.849	25.77	0.894
Retinexformer [15]	27.18	0.850	27.71	0.856	29.04	0.939
LAformer-T	**27.28**	0.868	**27.92**	0.868	**30.28**	**0.952**

demonstrates the efficiency and scalability of LAformer in both diffusion-based and flow-based visual generation. Fig. 8 shows that our method is capable of generating high-quality images efficiently.

4.4 Ablation Study

As shown in Tab. 14 and Tab. 13, we perform detailed ablation studies on SOTS-Indoor [49] and GoPro [69] datasets to analyze the role of each component in LAformer. For fair comparisons, all models are designed with comparable complexity. FLOPs are computed on 256 × 256 images. More ablations and analysis are provided in the **Appendix**.

Rank Enhanced Linear Attention. As shown in Tab. 14, RELA significantly improves performance over vanilla linear attention, achieving a 1.89 dB PSNR gain on SOTS-Indoor with only a modest increase of 0.07M parameters and 0.76G FLOPs. We also compare RELA with the commonly used window self-attention [53] and transposed self-attention [113] in image restoration. With similar parameter counts and FLOPs, our RELA achieves a PSNR improvement of 0.41 dB and 0.65 dB, respectively. This demonstrates the global perception capability of RELA, which contributes to improved restoration performance.

Table 10: All-in-one image restoration results for five tasks. We strictly follow the setting of DiffUIR [125].

Method	Venue	Deraining (5*sets*)		Enhancement		Desnowing (2*sets*)		Dehazing		Deblurring	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Restormer [113]	CVPR'22	27.10	0.843	17.63	0.542	28.61	0.876	22.79	0.706	26.36	0.814
AirNet [50]	CVPR'22	24.87	0.773	14.83	0.767	27.63	0.860	25.47	0.923	26.92	0.811
Painter [96]	CVPR'23	29.49	0.868	22.40	0.872	N/A	N/A	N/A	N/A	N/A	N/A
IDR [117]	CVPR'23	N/A	N/A	21.34	0.826	N/A	N/A	25.24	0.943	27.87	0.846
ProRes [62]	arXiv'23	30.67	0.891	22.73	0.877	N/A	N/A	N/A	N/A	27.53	0.851
PromptIR [72]	NeurIPS'23	29.56	0.888	22.89	0.847	31.98	0.924	32.02	0.952	27.21	0.817
DACLIP [61]	ICLR'24	28.96	0.853	24.17	0.882	30.80	0.888	31.39	0.983	25.39	0.805
DiffUIR [125]	CVPR'24	31.03	0.904	25.12	0.907	32.65	0.927	32.94	0.956	29.17	0.864
Varformer [94]	CVPR'25	31.33	0.913	25.13	0.917	N/A	N/A	32.96	0.956	N/A	N/A
LAformer-B	-	**31.55**	**0.914**	**26.81**	**0.924**	**34.24**	**0.942**	**34.28**	**0.984**	**30.92**	**0.910**

Table 11: Diffusion-based image generation results for 400K training steps on class-conditional ImageNet 256×256. We use *the same training hyperparameters* as DiT, DiC, and DiG.

Method	Venue	Params	FLOPs	FID↓	IS↑	Prec.↑	Rec.↑
DiT-S/2 [71]	ICCV'23	33M	6.06G	68.40	-	-	-
DiC-S [86]	CVPR'25	33M	5.9G	58.68	25.82	-	-
DiG-S/2 [128]	CVPR'25	33M	4.30G	62.06	22.81	0.39	0.56
LAformer	-	33M	4.32G	**50.39**	**28.64**	**0.44**	**0.59**
DiT-B/2	ICCV'23	130M	23.02G	43.47	-	-	-
DiC-B	CVPR'25	130M	23.5G	32.33	48.72	-	-
DiG-B/2	CVPR'25	130M	17.07G	39.50	37.21	0.51	0.63
LAformer	-	131M	17.32G	**26.92**	**53.17**	**0.56**	**0.64**
DiT-L/2	ICCV'23	458M	80.73G	23.33	-	-	-
DiG-L/2	CVPR'25	458M	61.66G	22.90	59.87	0.60	0.64
LAformer	-	463M	60.98G	**13.42**	**85.52**	**0.66**	**0.65**
DiT-XL/2	ICCV'23	675M	118.66G	19.47	-	-	-
DiC-XL	CVPR'25	702M	116.1G	13.11	**100.15**	-	-
DiG-XL/2	CVPR'25	676M	89.40G	18.53	68.53	0.63	0.64
LAformer	-	682M	89.91G	**11.31**	98.21	**0.67**	**0.66**

We also conduct ablations on the design of RELA, as shown in Tab. 13. We analyze the choice of activation function $\psi(\cdot)$, with experimental results indicating minimal impact on PSNR performance. This suggests that the effectiveness of our RELA is not dependent on meticulously chosen activation functions. We further analyze the impact of the depth-wise convolution (DWC) kernel size on the results. Considering both computational efficiency and performance, we ultimately select a 5×5 kernel.

Channel Attention Block. Tab. 14 shows that removing CAB results in a 0.36 dB drop in PSNR, highlighting the importance of channel-wise global information for IR.

Convolutional Gated FFN. Tab. 14 indicates that the DWC in CG-FFN significantly enhances LAformer's ability to extract local information, markedly

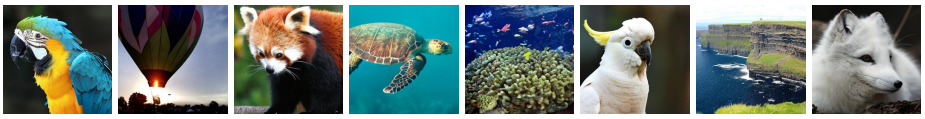

Fig. 8: Generated samples from LAformer trained on the class-conditional ImageNet dataset at 256×256 resolution.

Table 12: Flow-based image generation results for 400K steps. We use *the same training hyperparameters* as SiT and FlowDCN.

Method	Venue	Params	FLOPs	FID↓	IS↑	Prec↑	Rec↑
SiT-S/2 [64]	ECCV'24	33M	6.06G	57.64	24.78	0.41	0.60
FlowDCN-S/2 [93]	NeurIPS'24	30.3M	4.36G	54.6	26.4	-	-
LAformer	-	33M	4.32G	**46.92**	**31.74**	**0.46**	**0.61**
SiT-B/2	ECCV'24	130M	23.02G	33.02	43.71	0.53	0.63
FlowDCN-B/2	NeurIPS'24	130M	17.87G	28.5	51	-	-
LAformer	-	131M	17.32G	**24.37**	**55.53**	**0.57**	**0.64**
SiT-L/2	ECCV'24	458M	80.73G	18.79	72.02	0.64	0.64
FlowDCN-L/2	NeurIPS'24	421M	63.51G	13.8	85	-	-
LAformer	-	463M	60.98G	**12.38**	**87.34**	**0.66**	**0.65**
SiT-XL/2	ECCV'24	675M	118.66G	17.19	76.52	0.65	0.63
FlowDCN-XL/2	NeurIPS'24	618M	93.24G	11.3	97	-	-
LAformer	-	682M	89.91G	**10.98**	**102.04**	**0.67**	**0.66**

Table 13: Ablations on GoPro [69] for RELA designs.

RELA	Params (M)	FLOPs (G)	PSNR (dB)
$\psi(\cdot) = \text{ReLU}(\cdot)$	24.89	144.33	33.38
$\psi(\cdot) = 1 + \text{ELU}(\cdot)$	24.89	144.33	33.40
3×3 DWC	24.79	143.11	33.35
5×5 DWC	24.89	144.33	33.40
7×7 DWC	25.03	146.16	33.41

improving model performance. Removing it results in nearly a 1 dB drop in PSNR.

5 Conclusion

In this paper, we introduce linear attention to image restoration, aiming to overcome the complexity limitations of Transformer. We uncover the inherent low-rank limitations of vanilla linear attention and propose a simple yet effective solution: Rank Enhanced Linear Attention (RELA). RELA enhances the rank of features through a lightweight depth-wise convolution, thereby improving its representation capability. Building upon RELA, we introduce LAformer, a Vision Transformer designed to achieve effective global perception through the integration of linear attention and channel attention. Comprehensive experiments show

Table 14: Ablations on SOTS-Indoor [49] for LAformer-T.

Method	Params (M)	FLOPs (G)	PSNR (dB)
Vanilla LA [42]	6.18	29.08	39.28
Window SA [53]	6.20	30.19	40.76
Transposed SA [113]	6.26	29.87	40.52
Rank Enhanced LA	6.25	29.84	41.17
w/o CAB	6.13	30.79	40.81
Dual Attention	6.25	29.84	41.17
MLP	6.59	30.96	40.20
CG-FFN w/o DWC	6.24	30.16	40.11
CG-FFN w/o GLU	6.28	29.60	40.86
CG-FFN	6.25	29.84	41.17

that LAformer surpasses SOTA IR methods in performance, while also offering considerable computational efficiency gains. Furthermore, we extend LAformer to visual generation, showcasing its excellent efficiency and scalability. We aim to further scale up LAformer and extend it to broader generative tasks, including text-to-image generation.

References

1. Abuolaim, A., Brown, M.S.: Defocus deblurring using dual-pixel data. In: ECCV (2020) 9, 10, 11
2. Abuolaim, A., Delbracio, M., Kelly, D., Brown, M.S., Milanfar, P.: Learning to reduce defocus blur by realistically modeling dual-pixel data. In: ICCV (2021) 11
3. Agarap, A.: Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375 (2018) 2
4. Ai, Y., Cao, J., He, R., Huang, H.: Uncertainty-aware source-free adaptive image restoration with state space augmentation. IJCV (2026) 4
5. Ai, Y., Fan, Q., Hu, X., Yang, Z., He, R., Huang, H.: Dico: Revitalizing convnets for scalable and efficient diffusion modeling. In: NeurIPS (2025) 11
6. Ai, Y., Han, J., Zhuang, S., Mao, W., Hu, X., Yang, Z., Yang, Z., Wang, Y., Huang, H., Yue, X., et al.: Bitdance: Scaling autoregressive generative models with binary tokens. arXiv preprint arXiv:2602.14041 (2026) 4
7. Ai, Y., Huang, H., He, R.: Lora-ir: Taming low-rank experts for efficient all-in-one image restoration. arXiv preprint arXiv:2410.15385 (2024) 11
8. Ai, Y., Huang, H., Zhou, X., Wang, J., He, R.: Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In: CVPR (2024) 11
9. Ai, Y., Zhou, X., Huang, H., Han, X., Chen, Z., You, Q., Yang, H.: Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. In: NeurIPS (2024) 2
10. Ai, Y., Zhou, X., Huang, H., Zhang, L., He, R.: Uncertainty-aware source-free adaptive image super-resolution with wavelet augmentation transformer. In: CVPR (2024) 4
11. Ancuti, C.O., Ancuti, C., Sbert, M., Timofte, R.: Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In: ICIP (2019) 11, 12

12. Ancuti, C.O., Ancuti, C., Timofte, R., De Vleeschouwer, C.: O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In: CVPRW (2018) 11
13. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Hoffman, J.: Hydra attention: Efficient attention with many heads. In: ECCV (2022) 4
14. Cai, H., Li, J., Hu, M., Gan, C., Han, S.: Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In: ICCV (2023) 4, 8
15. Cai, Y., Bian, H., Lin, J., Wang, H., Timofte, R., Zhang, Y.: Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In: ICCV (2023) 12
16. Cao, M., Mou, C., Yu, F., Wang, X., Zheng, Y., Zhang, J., Dong, C., Li, G., Shan, Y., Timofte, R., et al.: Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results. In: CVPRW (2023) 4
17. Chen, C., Chen, Q., Do, M.N., Koltun, V.: Seeing motion in the dark. In: ICCV (2019) 12
18. Chen, H., Wang, Y., Guo, T., Xu, C., Deng, Y., Liu, Z., Ma, S., Xu, C., Xu, C., Gao, W.: Pre-trained image processing transformer. In: CVPR (2021) 4, 10, 12
19. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: ECCV (2022) 7, 10
20. Chen, W.T., Fang, H.Y., Ding, J.J., Tsai, C.C., Kuo, S.Y.: Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: ECCV (2020) 10
21. Chen, W.T., Fang, H.Y., Hsieh, C.L., Tsai, C.C., Chen, I., Ding, J.J., Kuo, S.Y., et al.: All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: ICCV (2021) 10
22. Chen, X., Li, H., Li, M., Pan, J.: Learning a sparse transformer network for effective image deraining. In: CVPR (2023) 2
23. Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C.: Activating more pixels in image super-resolution transformer. In: CVPR (2023) 2, 4
24. Chen, Z., Zhang, Y., Gu, J., Kong, L., Yang, X., Yu, F.: Dual aggregation transformer for image super-resolution. In: ICCV (2023) 2
25. Chen, Z., Zhang, Y., Liu, D., Gu, J., Kong, L., Yuan, X., et al.: Hierarchical integration diffusion model for realistic image deblurring. In: NeurIPS (2023) 10
26. Cho, S.J., Ji, S.W., Hong, J.P., Jung, S.W., Ko, S.J.: Rethinking coarse-to-fine approach in single image deblurring. In: ICCV (2021) 9, 10
27. Clevert, D.A.: Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 (2015) 2
28. Cui, Y., Ren, W., Cao, X., Knoll, A.: Focal network for image restoration. In: ICCV (2023) 10, 11
29. Cui, Y., Ren, W., Yang, S., Cao, X., Knoll, A.: Irnext: Rethinking convolutional network design for image restoration. In: ICML (2023) 10
30. Dauphin, Y.N., Fan, A., Auli, M., Grangier, D.: Language modeling with gated convolutional networks. In: ICLR (2017) 8
31. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 11
32. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. TPAMI **38**(2) (2015) 1, 4
33. Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., Yang, M.H.: Multi-scale boosted dehazing network with dense feature fusion. In: CVPR (2020) 11
34. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 4

35. Engin, D., Genç, A., Kemal Ekenel, H.: Cycle-dehaze: Enhanced cyclegan for single image dehazing. In: CVPRW (2018) 10
36. Fan, Q., Ai, Y., Huang, H., He, R.: Random wins all: Rethinking grouping strategies for vision tokens. In: CVPR (2026) 4
37. Fan, Q., Huang, H., Ai, Y., He, R.: Rectifying magnitude neglect in linear attention. In: ICCV (2025) 4
38. Greenspan, H.: Super-resolution in medical imaging. The computer journal **52**(1) (2009) 1
39. Gu, Y., Meng, Y., Ji, J., Sun, X.: Acl: Activating capability of linear attention for image restoration. In: CVPR) (2025) 10
40. Guo, C.L., Yan, Q., Anwar, S., Cong, R., Ren, W., Li, C.: Image dehazing transformer with transmission-aware 3d position embedding. In: CVPR (2022) 11, 12
41. Han, D., Pan, X., Han, Y., Song, S., Huang, G.: Flatten transformer: Vision transformer using focused linear attention. In: ICCV (2023) 4
42. Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F.: Transformers are rnns: Fast autoregressive transformers with linear attention. In: ICML (2020) 2, 4, 5, 7, 15
43. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016) 4
44. Kim, K., Lee, S., Cho, S.: Mssnet: Multi-scale-stage network for single image deblurring. In: ECCV (2022) 9
45. Kong, L., Dong, J., Ge, J., Li, M., Pan, J.: Efficient frequency domain-based transformers for high-quality image deblurring. In: CVPR (2023) 9
46. Kupyn, O., Martyniuk, T., Wu, J., Wang, Z.: Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In: ICCV (2019) 9
47. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017) 4
48. Lee, J., Son, H., Rim, J., Cho, S., Lee, S.: Iterative filter adaptive network for single image defocus deblurring. In: CVPR (2021) 11
49. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. TIP **28**(1) (2018) 11, 12, 15
50. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: CVPR (2022) 13
51. Li, R., Tan, R.T., Cheong, L.F.: All in one bad weather removal using architectural search. In: CVPR (2020) 10
52. Li, Y., Fan, Y., Xiang, X., Demandolx, D., Ranjan, R., Timofte, R., Van Gool, L.: Efficient and explicit modelling of image hierarchies for image restoration. In: CVPR (2023) 2, 4, 6, 9
53. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: ICCVW (2021) 2, 4, 12, 15
54. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: CVPRW (2017) 4
55. Liu, R., Ma, L., Zhang, J., Fan, X., Luo, Z.: Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In: CVPR (2021) 12
56. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: Attention-based multi-scale network for image dehazing. In: ICCV (2019) 11
57. Liu, X., Suganuma, M., Sun, Z., Okatani, T.: Dual residual networks leveraging the potential of paired operations for image restoration. In: CVPR (2019) 12

58. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. TIP **27**(6) (2018) 10
59. Liu, Y., Huang, T., Dong, W., Wu, F., Li, X., Shi, G.: Low-light image enhancement with multi-stage residue quantization and brightness-aware attention. In: ICCV (2023) 12
60. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021) 4
61. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Controlling vision-language models for universal image restoration. arXiv preprint arXiv:2310.01018 (2023) 13
62. Ma, J., Cheng, T., Wang, G., Zhang, Q., Wang, X., Zhang, L.: Prores: Exploring degradation-aware visual prompt for universal image restoration. arXiv preprint arXiv:2306.13653 (2023) 13
63. Ma, L., Ma, T., Liu, R., Fan, X., Luo, Z.: Toward fast, flexible, and robust low-light image enhancement. In: CVPR (2022) 12
64. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: ECCV (2024) 11, 14
65. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: CVPR (2017) 4
66. Mao, X., Liu, Y., Shen, W., Li, Q., Wang, Y.: Deep residual fourier transformation for single image deblurring. arXiv preprint arXiv:2111.11745 (2021) 9
67. Mei, Y., Fan, Y., Zhou, Y.: Image super-resolution with non-local sparse attention. In: CVPR (2021) 4
68. Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T.S., Shi, H.: Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In: CVPR (2020) 4
69. Nah, S., Hyun Kim, T., Mu Lee, K.: Deep multi-scale convolutional neural network for dynamic scene deblurring. In: CVPR (2017) 9, 10, 12, 14
70. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. TPAMI **45**(8) (2023) 12
71. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023) 11, 13
72. Potlapalli, V., Zamir, S.W., Khan, S., Khan, F.S.: Promptir: Prompting for all-in-one blind image restoration. arXiv preprint arXiv:2306.13090 (2023) 13
73. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: CVPR (2018) 10, 12
74. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: AAAI (2020) 11
75. Qiu, Y., Zhang, K., Wang, C., Luo, W., Li, H., Jin, Z.: Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In: ICCV (2023) 11, 12
76. Quan, Y., Deng, S., Chen, Y., Ji, H.: Deep learning for seeing through window with raindrops. In: ICCV (2019) 12
77. Quan, Y., Wu, Z., Ji, H.: Neumann network with recursive kernels for single image defocus deblurring. In: CVPR (2023) 9, 11
78. Quan, Y., Yao, X., Ji, H.: Single image defocus deblurring via implicit neural inverse kernels. In: ICCV (2023) 11

79. Rasti, P., Uiboupin, T., Escalera, S., Anbarjafari, G.: Convolutional neural network super resolution for face recognition in surveillance monitoring. In: Articulated Motion and Deformable Objects (2016) 1
80. Rim, J., Lee, H., Won, J., Cho, S.: Real-world blur dataset for learning and benchmarking deblurring algorithms. In: ECCV (2020) 9
81. Ruan, L., Chen, B., Li, J., Lam, M.: Learning to deblur using light field generated and real defocus images. In: CVPR (2022) 9, 11
82. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: WACV (2021) 4
83. Shen, Z., Wang, W., Lu, X., Shen, J., Ling, H., Xu, T., Shao, L.: Human-aware motion deblurring. In: ICCV (2019) 9, 10
84. Sun, S., Ren, W., Gao, X., Wang, R., Cao, X.: Restoring images in adverse weather conditions via histogram transformer. In: ECCV (2024) 10, 12
85. Tao, X., Gao, H., Shen, X., Wang, J., Jia, J.: Scale-recurrent network for deep image deblurring. In: CVPR (2018) 9
86. Tian, Y., Han, J., Wang, C., Liang, Y., Xu, C., Chen, H.: Dic: Rethinking conv3x3 designs in diffusion models. In: CVPR (2025) 13
87. Tsai, F.J., Peng, Y.T., Lin, Y.Y., Tsai, C.C., Lin, C.W.: Stripformer: Strip transformer for fast image deblurring. In: ECCV (2022) 9, 10
88. Tsai, F.J., Peng, Y.T., Tsai, C.C., Lin, Y.Y., Lin, C.W.: Banet: a blur-aware attention network for dynamic scene deblurring. TIP **31** (2022) 9
89. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxim: Multi-axis mlp for image processing. In: CVPR (2022) 11, 12
90. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: CVPR (2022) 10
91. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 2, 4, 5
92. Wang, C., Pan, J., Wang, W., Dong, J., Wang, M., Ju, Y., Chen, J.: Promptrestorer: A prompting image restoration method with degradation perception. In: NeurIPS (2023) 10
93. Wang, S., Li, Z., Song, T., Li, X., Ge, T., Zheng, B., Wang, L.: Flowdcn: Exploring dcn-like architectures for fast image generation with arbitrary resolution. In: NeurIPS (2024) 14
94. Wang, S., Zheng, N., Huang, J., Zhao, F.: Navigating image restoration with var's distribution alignment prior. In: CVPR (2025) 13
95. Wang, T., Xia, P., Li, B., Jiang, P.T., Kong, Z., Zhang, K., Lu, T., Luo, W.: Moerl: When mixture-of-experts meet reinforcement learning for adverse weather image restoration. In: ICCV (2025) 12
96. Wang, X., Wang, W., Cao, Y., Shen, C., Huang, T.: Images speak in images: A generalist painter for in-context visual learning. In: CVPR (2023) 13
97. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y., Tang, X.: Esrgan: Enhanced super-resolution generative adversarial networks. arXiv preprint arXiv:1809.00219 (2018) 4
98. Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.P., Kot, A.: Low-light image enhancement with normalizing flow. In: AAAI (2022) 12
99. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: CVPR (2022) 2, 4, 7, 10, 11, 12
100. Wei, C., Wang, W., Yang, W., Liu, J.: Deep retinex decomposition for low-light enhancement. arXiv preprint arXiv:1808.04560 (2018) 12

101. Wu, H., Qu, Y., Lin, S., Zhou, J., Qiao, R., Zhang, Z., Xie, Y., Ma, L.: Contrastive learning for compact single image dehazing. In: CVPR (2021) 11
102. Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., Jiang, J.: Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In: CVPR (2022) 12
103. Xia, B., Hang, Y., Tian, Y., Yang, W., Liao, Q., Zhou, J.: Efficient non-local contrastive attention for image super-resolution. In: AAAI (2022) 4
104. Xia, B., Zhang, Y., Wang, S., Wang, Y., Wu, X., Tian, Y., Yang, W., Van Gool, L.: Diffir: Efficient diffusion model for image restoration. In: ICCV (2023) 3
105. Xiao, J., Fu, X., Liu, A., Wu, F., Zha, Z.J.: Image de-raining transformer. TPAMI **45**(11) (2023) 12
106. Xu, L., Ren, J.S., Liu, C., Jia, J.: Deep convolutional neural network for image deconvolution. In: NeurIPS (2014) 1
107. Xu, X., Wang, R., Fu, C.W., Jia, J.: Snr-aware low-light image enhancement. In: CVPR (2022) 12
108. Yang, W., Wang, S., Fang, Y., Wang, Y., Liu, J.: From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement. In: CVPR (2020) 12
109. Yang, W., Wang, W., Huang, H., Wang, S., Liu, J.: Sparse gradient regularized deep retinex network for robust low-light image enhancement. TIP **30** (2021) 12
110. Ye, T., Chen, S., Bai, J., Shi, J., Xue, C., Jiang, J., Yin, J., Chen, E., Liu, Y.: Adverse weather removal with codebook priors. In: ICCV (2023) 12
111. Ye, T., Chen, S., Chai, W., Xing, Z., Qin, J., Lin, G., Zhu, L.: Learning diffusion texture priors for image restoration. In: CVPR (2024) 12
112. Ye, T., Zhang, Y., Jiang, M., Chen, L., Liu, Y., Chen, S., Chen, E.: Perceiving and modeling density for image dehazing. In: ECCV (2022) 11
113. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: CVPR (2022) 2, 3, 4, 7, 9, 10, 11, 12, 13, 15
114. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Learning enriched features for real image restoration and enhancement. In: ECCV (2020) 12
115. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021) 9, 10
116. Zhang, J., Zhang, Y., Gu, J., Zhang, Y., Kong, L., Yuan, X.: Accurate image restoration with attention retractable transformer. In: ICLR (2023) 2, 4
117. Zhang, J., Huang, J., Yao, M., Yang, Z., Yu, H., Zhou, M., Zhao, F.: Ingredient-oriented multi-degradation learning for image restoration. In: CVPR (2023) 13
118. Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L.: Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. TIP **26**(7) (2017) 1, 4
119. Zhang, X., Zhang, Y., Yu, F.: Hit-sr: Hierarchical transformer for efficient image super-resolution. In: ECCV (2024) 4
120. Zhang, Y., Zhang, J., Guo, X.: Kindling the darkness: A practical low-light image enhancer. In: ACMMM (2019) 12
121. Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y.: Image super-resolution using very deep residual channel attention networks. In: ECCV (2018) 3, 4, 8
122. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018) 4
123. Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image restoration. TPAMI **43**(7) (2021) 4

124. Zhang, Y., Zhang, K., Chen, Z., Li, Y., Timofte, R., Zhang, J., Zhang, K., Peng, R., Ma, Y., Jia, L., et al.: Ntire 2023 challenge on image super-resolution (x4): Methods and results. In: CVPRW (2023) 4
125. Zheng, D., Wu, X.M., Yang, S., Zhang, J., Hu, J.F., Zheng, W.S.: Selective hourglass mapping for universal image restoration based on diffusion model. In: CVPR (2024) 11, 13
126. Zhou, S., Chen, D., Pan, J., Shi, J., Yang, J.: Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In: CVPR (2024) 2, 4, 10, 11, 12
127. Zhou, S., Pan, J., Shi, J., Chen, D., Qu, L., Yang, J.: Seeing the unseen: A frequency prompt guided transformer for image restoration. In: ECCV (2024) 9, 10
128. Zhu, L., Huang, Z., Liao, B., Liew, J.H., Yan, H., Feng, J., Wang, X.: Dig: Scalable and efficient diffusion models with gated linear attention. In: CVPR (2025) 13
129. Zhu, Z., Liang, D., Zhang, S., Huang, X., Li, B., Hu, S.: Traffic-sign detection and classification in the wild. In: CVPR (2016) 1, 4