ACL: Activating Capability of Linear Attention for Image Restoration

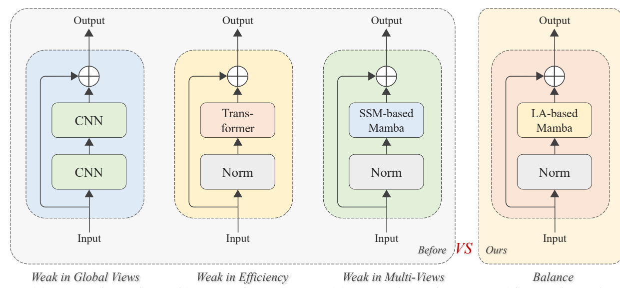
Yubin Gu[1], Yuan Meng[1], Jiayi Ji[1,2], Xiaoshuai Sun[1]*
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China
[2] National University of Singapore, Singapore
{guyubin,aprilmyy}@stu.xmu.edu.cn, jjyxmu@gmail.com, xssun@xmu.edu.cn

Abstract

Image restoration (IR), a key area in computer vision, has entered a new era with deep learning. Recent research has shifted toward Selective State Space Models (Mamba) to overcome CNNs' limited receptive fields and Transformers' computational inefficiency. However, due to Mamba's inherent one-dimensional scanning limitations, recent approaches have introduced multi-directional scanning to bolster inter-sequence correlations. Despite these enhancements, these methods still struggle with managing local pixel correlations across various directions. Moreover, the recursive computation in Mamba's SSM leads to reduced efficiency. To resolve these issues, we exploit the mathematical congruences between linear attention and SSM within Mamba to propose a novel model, ACL, which leverages news designs to Activate the Capability of Linear attention for IR. ACL integrates linear attention blocks instead of SSM within Mamba, serving as the core component of encoders/decoders, and aims to preserve a global perspective while boosting computational efficiency. Furthermore, we have designed a simple yet robust local enhancement module with multi-scale dilated convolutions to extract both coarse and fine features to improve local detail recovery. Experimental results confirm that our ACL model excels in classical IR tasks such as de-raining and de-blurring, while maintaining relatively low parameter counts and FLOPs[1].

1. Introduction

In the field of image processing, restoring degraded images to clarity is a crucial technology. High-quality image restoration (IR) methods play a key role in ensuring the smooth progression of downstream vision tasks. Vanilla IR methods rely primarily on manually designed feature extraction, but often perform poorly when faced with complicated degradation factors in reality.



(a) CNN-based Block (b) Transformer-based Block (c) SSM-based Mamba Block (d) LA-based Mamba Block

Figure 1. Comparison of our method's core block design with those of recent mainstream approaches.

Over the past decade, with the rapid development of deep learning, many fields have been propelled forward, such as image segmentation and generation [11, 12, 14, 17, 48], and of course, IR as well. The initial IR models were primarily based on CNN designs [6, 41], whose translational invariance and high inferential efficiency have made them widely applied, as depicted in the core structure in Fig. 1(a). However, these CNN-based models face challenges in processing global features and often require stacking additional network layers to compensate for this deficiency. With the rise of attention mechanisms [34], particularly the application of the Transformer, which boasts exceptional global feature modeling capabilities, an increasing number of researchers have shifted toward Transformer-based designs [23, 36, 44], achieving remarkable restoration results, as shown in the core structure in Fig. 1 (b). Although the Transformer broadens the global perspective, its softmax attention mechanism's quadratic computational complexity significantly reduces efficiency during inference.

Recently, state-space models from the field of control science, especially the Mamba model [8], have been proposed by related IR methods [13, 30] demonstrating superior training and inference efficiency on sequential data. These models have even outperformed some Transformer-

*Corresponding Author: Xiaoshuai Sun
[1]https://github.com/ClimBin/ACLNet

based IR methods in terms of performance. Mamba is initially designed for one-dimensional sequence data modeling, direct adaptation to two-dimensional image poses challenges. The improved SSM module in Mamba, a one-dimensional unidirectional scanning, recursive computing structure, faces major issues when applied to the image, including: 1) conversion of image data into a one-dimensional sequence increases the distance between adjacent pixels, leading to the loss of local relationships [30]; 2) unidirectional modeling neglects the spatial relationships of pixels in multiple directions. Recently, MambaIR [13] and VMambaIR [30] have adopted multi-directional scanning methods to mitigate these problems, as illustrated in the main structure in Fig. 1 (c), but these methods have yet to effectively and directly establish multi-directional connections in the spatial dimension of pixels.

In response to the limitations of the Mamba model, we propose a method that aims to leverage the advantages of the Mamba structure while addressing its unidirectional modeling constraint. A direct approach to overcoming this limitation is to introduce a long-range dependency attention mechanism, such as Linear Attention (LA) or Softmax Attention (SA). Although SA generally outperforms LA in traditional vision tasks, its computational complexity is significantly higher than that of LA. Upon further analysis, we observe that LA and the State-Space Model (SSM) in Mamba share a highly similar mathematical formulation, which has also been deeply analyzed in the work [15]. Inspired by this insight, we replace the SSM module in Mamba with LA layers, thereby designing a novel foundational IR architecture, as illustrated in Fig. 1 (d), and propose a new IR model, ACL.

Specifically, the proposed ACL consists of two main components: the Mamba-based module with LA at its core (denoted as LAMA), which serves as the central part of the encoder/decoder and establishes global feature dependencies with linear computational complexity. Additionally, optimizing local features is equally essential. A recent approach involves partitioning the global features in the spatial domain into small windows and optimizing attention within these windows to enhance local feature modeling. While this approach has shown some effectiveness, it incurs high computational costs. In contrast, we propose a simple yet efficient multi-scale dilated convolution module (MDC), which captures local features at varying granularities by employing different dilation factors, thus improving detail restoration. By combining the Mamba structure with LA, ACL activates the capability of LA, achieving advanced performance on two classic IR tasks—deblurring and deraining. Compared to Transformer-based models [32, 44], ACL not only significantly reduces computational cost and parameter count, but also achieves superior or comparable performance.

Our contributions are summarized as follows:

- We explore an alternative CNN and Transformer-based architecture for Image Restoration, providing global receptive fields while maintaining computational efficiency. Specifically, we propose ACL, which upgrades the SSM in Mamba with the LA structure, enabling the model to perform global multi-directional scanning.
- The proposed ACL consists of two modules: LAMA and MDC. LAMA captures global feature dependencies, and MDC captures local features at varying granularities. Both modules enhance detail restoration.
- The proposed ACL demonstrates advanced performance in de-raining and de-blurring tasks, proving its advantages in terms of parameter count and computational cost.

2. Related Work

2.1. Image Restorations

Image restoration (IR) technology provides clear visual data essential for numerous advanced downstream visual tasks. In recent years, the advent of deep learning has eclipsed traditional methods that rely on manually designed features, shifting the paradigm toward deep-learning-based approaches [9, 13, 20, 22, 45]. Initially, models were predominantly designed using CNNs, achieving significant advancements through sophisticated network designs that incorporated encoder-decoder patterns [5], dense connections, and residual connections [10]. However, CNN-based techniques face challenges in establishing global feature dependencies. With the evolution of Transformers in both Natural Language Processing (NLP) and Computer Vision (CV), many researchers have pivoted towards Transformer-based IR methods [23, 38, 44], leveraging their robust global receptive capabilities and marking substantial progress. Despite these improvements, the computational burden of calculating SA remains a drawback. Recently, the Mamba model [8], known for its efficient training and inference capabilities, has introduced a new potential paradigm in the IR field. Methods such as VMambaIR [30] and MambaIR [13], which employ multi-directional scanning strategies, aim to address the issues of unidirectional modeling inherent in Mamba's SSM blocks. Nonetheless, these methods have yet to establish multi-directional pixel connections directly. Thus, exploring how to utilize Mamba's superior design to establish comprehensive multidirectional global pixel correlations remains a promising direction.

2.2. Attention Mechanisms

Originally applied in the NLP field, attention mechanisms [34], particularly the Transformer with its SA mechanism, have achieved remarkable success. Subsequently, these mechanisms have been successfully adapted to the visual domain. In recent years, numerous IR methods based on the Transformer structure have emerged [1, 3, 23, 24], primarily utilizing various design modules to leverage self-attention mechanisms and enhance model efficiency. SwinIR [23] introduces twin-shifted window attention to boost performance. IPT [1] enhances local detail restoration by dividing images into multiple small windows and processing each window's features independently. However, adopting SA to establish global or local feature dependencies inevitably leads to quadratic computational complexity. Linear attention, which operates with linear complexity, has yet to match the performance of SA in classical vision tasks, thus its application remains limited. In this paper, we introduce a new IR model, ACL, activating the potential of linear attention.

2.3. State Space Model

The Mamba model, a newly proposed state space model (SSM), effectively facilitates sequence modeling with linear complexity [8]. Owing to its efficient training and inference speeds, many researchers have adapted it for visual tasks [16, 25, 30, 39, 50]. For instance, LocalMamba [16] employs a cross-scanning module to scan image spaces. VMambaIR [30] enhances multidirectional relationships between pixels by scanning images from six directions. These methods focus on overcoming the limitations of unidirectional modeling in SSMs by proposing various scanning techniques. However, the features captured in each direction remain unidirectionally connected and potentially lead to redundant computational costs. The Mamba model is efficient due to its structural design, yet its unidirectional scanning is not entirely suitable for images. Therefore, we combine linear attention with the Mamba structure to achieve a new balance between computational efficiency and restoration effectiveness.

3. Methods

3.1. Preliminary Analysis

In recent years, very few IR models based on linear attention have been proposed, as they tend to perform slightly worse than SA in classical vision tasks. Leveraging the computational advantages of Linear Attention (LA) to further explore its potential in visual tasks is crucial.

The improved SSM in Mamba shows significant potential in sequence processing. In fact, linear attention has a similar expression to SSM [15]. In linear attention, if the attention of the i-th token is restricted to only be related to the previous i tokens, it is expressed as follows:

$$\mathbf{A}_i = \frac{\mathbf{Q}_i\left(\sum_{j=1}^{i}\mathbf{K}_j^\top\mathbf{V}_j\right)}{\mathbf{Q}_i\left(\sum_{j=1}^{i}\mathbf{K}_j^\top\right)} = \frac{(\mathbf{Q}_i\mathbf{D}_i)}{(\mathbf{Q}_i\mathbf{U}_i)}, \tag{1}$$

where $\mathbf{D}_i = \sum_{j=1}^{i}\mathbf{K}_j^\top\mathbf{V}_j$, $\mathbf{U}_i = \sum_{j=1}^{i}\mathbf{K}_j^\top$. Therefore, the recursive expressions are:

$$\mathbf{U}_i = \mathbf{U}_{i-1} + \mathbf{K}_i^\top, \tag{2}$$

$$\mathbf{D}_i = \mathbf{D}_{i-1} + \mathbf{K}_i^\top\mathbf{V}_i, \quad \mathbf{A}_i = \frac{(\mathbf{Q}_i\mathbf{D}_i)}{(\mathbf{Q}_i\mathbf{U}_i)}. \tag{3}$$

To enable applications in deep neural networks, it is necessary to discretize the initial SSM using zero-order hold [8]. This involves discretizing the continuous parameters $\mathbf{A}$ and $\mathbf{B}$ into $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$ through the time scale parameter Δ. The specific expressions are as follows:

$$\mathbf{h}_i = \bar{\mathbf{A}}\mathbf{h}_{i-1} + \mathbf{B}\mathbf{x}_i, \quad \mathbf{y}_i = \mathbf{C}\mathbf{h}_i + \mathbf{D}\mathbf{x}_i, \tag{4}$$

where $\bar{\mathbf{A}} = \exp(\Delta\mathbf{A})$ and $\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - I)\Delta\mathbf{B} \approx \Delta\mathbf{B}$. For simplicity, we have omitted the feature dimension information of each part in the formulas.

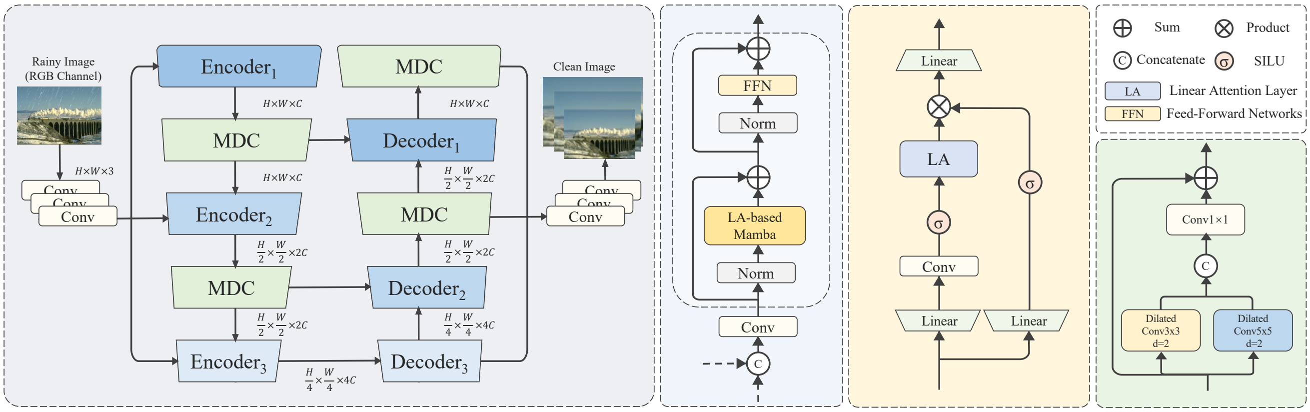
Further, Mamba enhances the discrete SSM by making $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, and Δ dependent on the input $\mathbf{x}_i$, breaking away from the assumption of input-independent models. Additionally, since $\bar{\mathbf{A}}_i$ in Mamba is a diagonal matrix, we have $\tilde{\mathbf{A}}_i = diag(\bar{\mathbf{A}}_i)$. The expression is thus transformed into:

$$\mathbf{h}_i = \tilde{\mathbf{A}}_i\mathbf{h}_{i-1} + \mathbf{B}_i(\Delta_i\mathbf{x}_i), \quad \mathbf{y}_i = \mathbf{C}_i\mathbf{h}_i + \mathbf{D}\mathbf{x}_i. \tag{5}$$

The primary distinctions between Eq. 3 and Eq. 5 are as follows: 1) The improved State Space Model (SSM) incorporates an additional parameter for hidden state transitions, denoted as $\tilde{\mathbf{A}}_i$. This parameter functions similarly to a forget gate, filtering previous states to enhance selective retention. 2) An additional term, $\mathbf{D}\mathbf{x}_i$, is introduced, akin to an input skip connection. In the Mamba model, which aims to achieve input-dependent modeling, $\tilde{\mathbf{A}}_i$ must be recursively computed. Despite the utilization of hardware acceleration mechanisms, this process still adheres to unidirectional modeling. Linear attention represents an alternative form of SSM within Mamba and can transcend the limitations of unidirectional pixel modeling, presenting a potential capability. For a more in-depth analysis, one can refer to another outstanding analytical works [15].

3.2. Overall Structure of the Model

As illustrated in Fig. 2 (a), the proposed IR model, ACL, is based on an encoder-decoder architecture. In this model, multiple downsampled degraded images are fed into the main pathway of the encoder through lateral convolution layers, and images restored at three different scales are output during the decoding phase. Both the encoder and decoder comprise three fundamental core blocks, the structure of which is depicted in Fig. 2. Each core unit consists of several successive LAMA modules. Furthermore, as shown in the framework diagram, following two core units with higher feature resolution in both encoding and decoding



(a) Overall Pipeline (b) Encoder and Decoder (c) LA-based Mamba (d) Multi-Dilated Convolutions Module

Figure 2. (a) The overall framework of ACL, based on the encoder-decoder architecture. (b) The core structure of the encoder-decoder, which includes the improved LA-based Mamba (LAMA) module. (c) The structure of the LAMA module. (d) The MDC module.

phases, a local enhancement module is appended to augment the model's capability for local detail restoration. The structure of this local enhancement module is presented in Fig. 2 (d).

Specifically, the network process begins with a degraded image $\mathbf{I} \in \mathbb{R}^{3\times256\times256}$. The model first transforms this image through a convolution layer into a feature map $\mathbf{I}' \in \mathbb{R}^{C\times H\times W}$, expanding the number of feature channels from 3 to C, where $C = 32$. Subsequently, this feature map is further processed through three encoder units E_i and a local enhancement module, each unit encoding the feature maps at different scales into a latent space state, denoted as $\mathbf{I}_e^i \in \mathbb{R}^{2^{(i-1)}C\times\frac{H}{2^{(i-1)}}\times\frac{W}{2^{(i-1)}}}$, where $i = 1, 2, 3$. Here, C, H, and W respectively represent the number of feature channels, the height, and the width of the feature maps. These processed feature maps are then passed to the decoder, where they are fused through direct or skip connections. In the decoder, the feature maps are gradually restored by decoding units D_i, generating $\mathbf{I}_d^i \in \mathbb{R}^{2^{(i-1)}C\times\frac{H}{2^{(i-1)}}\times\frac{W}{2^{(i-1)}}}$. Each feature map is processed by the subsequent D_i and, following lateral convolution operations, yields multi-scale output results, with D_1 producing the final restored image. Subsequently, we will elaborate on the two key modules that constitute ACL and the model's optimization function.

3.3. Linear Attention-based Mamba Module

The original Mamba model is an auto-regressive model capable of efficiently capturing sequence dependencies, and it has been proven effective in modeling temporal causal sequence data. However, due to its unidirectional modeling approach, Mamba exhibits limitations when handling data with weak causality, such as images, necessitating further improvements to address these challenges. To overcome the limitations of unidirectional modeling, recent methods have proposed multi-directional cross-scanning techniques for image processing. Unlike these approaches, we embed linear attention into the Mamba structure, enabling it to establish global pixel dependencies when processing image data, and avoiding the need for recursive computation of the forget matrix $\bar{\mathbf{A}}_i$. The module structure is illustrated in Fig. 2 (c).

The input to the model is a feature map $\mathbf{F} \in \mathbb{R}^{B\times C\times H\times W}$. First, $\mathbf{F}$ undergoes a dimensional transformation, resulting in $\mathbf{F}' \in \mathbb{R}^{B\times HW\times C}$. Subsequently, $\mathbf{F}'$ is passed into two branches: a main branch and a residual branch. The operations of the residual branch can be expressed as follows:

$$\mathbf{F}_{res} = \sigma(\text{Linear}(\mathbf{F}')), \tag{6}$$

where $\sigma(\cdot)$ represents the SiLU activation function.

The main branch comprises a linear mapping layer, a convolutional layer, and linear attention. The process can be expressed as follows:

$$\mathbf{F}_1 = \text{To4D}(\text{Linear}(\mathbf{F}')), \tag{7}$$

$$\mathbf{F}_2 = \sigma(\text{Conv}(\mathbf{F}_1)), \tag{8}$$

where To4D$(\cdot)$ indicates reshaping the feature map into a four-dimensional tensor to adapt convolution operation. Next, linear attention is applied to $\mathbf{F}_2$. The specific process is expressed as follows:

$$\mathbf{F}_2' = \text{To3D}(\mathbf{F}_2), \tag{9}$$

$$\mathbf{Q} = \phi(\text{Linear}(\mathbf{F}_2')), \quad \mathbf{K} = \phi(\text{Linear}(\mathbf{F}_2')), \tag{10}$$

$$\mathbf{KV} = \mathbf{K}^\top \cdot \mathbf{F}_2', \tag{11}$$

$$\mathbf{F}_{atten} = \mathbf{Q} \cdot \mathbf{KV}, \tag{12}$$

$$\mathbf{F}_{atten} = \mathbf{F}_{atten} + \mathrm{Conv}_{pos}(\mathbf{F}_2), \tag{13}$$

where To3D$(\cdot)$ reshapes the feature map into a three-dimensional tensor for linear attention layer, $\phi(\cdot)$ is the kernel function, and Conv$_{pos}(\cdot)$ is learnable position embedding function. Subsequently, $\mathbf{F}_{atten}$ is multiplied by $\mathbf{F}_{res}$, followed by a linear mapping layer, yielding $\mathbf{F}_{atten}$, which is expressed as:

$$\mathbf{F}_{enhence} = \mathrm{Linear}(\mathbf{F}_{atten} \times \mathbf{F}_{res}). \tag{14}$$

Finally, $\mathbf{F}_{enhence}$ is processed through a simple feed-forward neural network to produce the output of the LAMA.

3.4. Multi-Dilated Convolutions Module

The LAMA module primarily functions to establish global feature connections, necessitating the learning of local features to enhance the quality of detail restoration. While some previous methods employed feature window-based self-attention strategies and achieved certain advancements, they incurred substantial computational costs. Consequently, we adopted a more straightforward and effective approach, namely the multi-scale dilated convolution module, the structure of which is depicted in Fig. 2(d). This module is equipped with filtering operations using various dilation factors aimed at capturing local features of different granularities within the image to enhance the detail restoration effects. The module comprises two dilated convolutions along with skip connections. The input feature, denoted as $\mathbf{F}$, is split into two pathways: one passes through a convolution layer with a kernel size of 5 and a dilation rate of 2, and the other through a convolution layer with a kernel size of 3 and a dilation rate of 2, resulting in two feature sets:

$$\mathbf{F}_1 = \mathrm{Conv}_{5\times5,d=2}(\mathbf{F}), \tag{15}$$

$$\mathbf{F}_2 = \mathrm{Conv}_{3\times3,d=2}(\mathbf{F}). \tag{16}$$

These are then concatenated to form $\mathbf{F}'$, which subsequently passes through a convolution layer with a kernel size of 1 to halve the channel count, aligning it with the dimensions of the input features. Furthermore, the input feature $\mathbf{F}$ is merged with $\mathbf{F}'$ via a skip connection, culminating in the output $\mathbf{F}_{out}$. The expressions are as follows:

$$\mathbf{F}_{out} = \mathrm{Conv}_{1\times1}(\mathrm{Concat}(\mathbf{F}_1, \mathbf{F}_2)) + \mathbf{F}. \tag{17}$$

3.5. Optimization Objectives

The ACL model, during its decoding phase, outputs restoration results at three distinct scales and computes the corresponding loss values. Following prior methodologies, we calculate the loss values concurrently in both the spatial and frequency domains. The traditional L1 loss function is employed to measure the discrepancy between the restored outputs and the pristine reference images. Consequently, the total loss is computed as follows:

$$L_{total} = \sum_{i=1}^{3} \frac{1}{N_i} \left|\mathbf{P}_i - \mathbf{I}_i\right| + \lambda \cdot \sum_{i=1}^{3} \frac{1}{N_i} \left|fft(\mathbf{P}_i) - fft(\mathbf{I}_i)\right| \tag{18}$$

where $\mathbf{P}_i$ and $\mathbf{I}_i$ represent the restored result and the corresponding true image, respectively, and N_i denotes the total number of pixels in the image. $fft(\cdot)$ signifies the fast Fourier transform function. The hyperparameter λ, utilized to balance the contributions of spatial domain loss and frequency domain loss to the total loss, is set at 0.1, following previous methods.

4. Experiments

This section focuses on showcasing the effectiveness of our proposed ACL model in addressing various degraded image tasks, such as deraining and deblurring, evaluated across six test sets. We will outline the experimental procedures and datasets utilized, and confirm the impact of the proposed modules through a series of ablation studies.

4.1. Implementation Setup

For each type of degradation, datasets are trained and evaluated independently. Unless specifically mentioned, all tests are conducted with the same hyperparameters. The training set undergoes random cropping of 256 $\times$ 256 patches and random flipping as a data augmentation strategy. To compare computational complexity with other methods, FLOPs are tested at the mentioned crop size. A cosine annealing strategy is adopted to gradually adjust the learning rate each epoch, setting a minimum learning rate limit of 1e-6. The batch size is set to 8, and the Adam optimizer is adopted. Following the evaluation of previous methods, for deraining, the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) are calculated on the YCbCr color mode. For other tasks, evaluation metrics are calculated on RGB color mode. All experiments are implemented in an environment equipped with the NVIDIA 24GB 3090 GPUs, based on the PyTorch.

4.2. Single Image Deraining Results

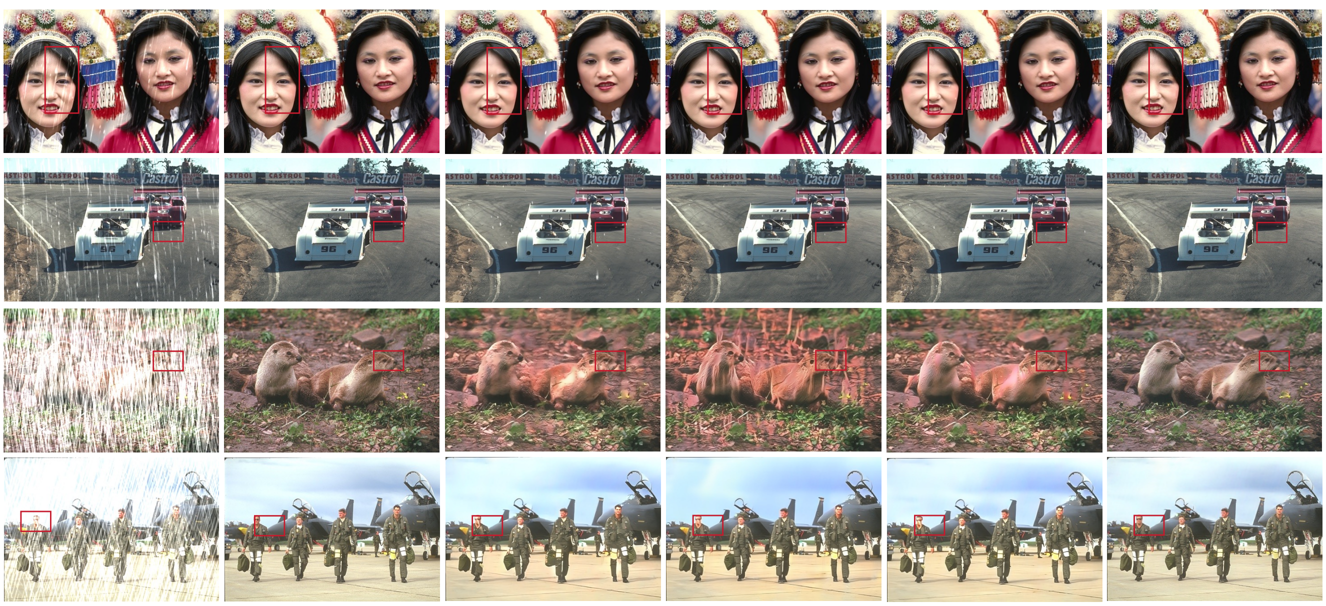
We utilized several different rain removal datasets, including Rain100L/H [40], Rain200L/H, and DID-Data [46], to evaluate the model's ability to restore images with streak-like degradation elements. Each dataset was independently trained for 800 epochs, with an initial learning rate set to 1e-3. We compared our approach with previous methods, including the advanced Restormer (Transformer) [44], NAFNet (CNN) [2], IRNeXt (CNN) [6], and MambaIR (Mamba) [13]. The comparative results are shown in Table 1 and Table 2. Upon comparison of PSNR, it can

Table 1. Quantitative comparison results of the proposed model and seven other advanced models on the Rain100L and Rain100H. The larger the PSNR and SSIM values, the better the model effect.

Methods	Restormer [44]		MAXIM [33]		DRT [24]		MPRNet [43]		DAWN [19]		IRNeXt [6]		MambaIR [13]		ACL(Ours)	
Dataset	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Rain100L	38.99	0.978	38.06	0.977	37.61	0.948	37.84	0.959	36.73	0.971	38.24	0.972	38.78	0.977	**39.18**	**0.983**
Rain100H	31.46	0.904	30.81	0.903	29.47	0.846	30.41	0.874	30.62	0.896	31.64	0.902	30.62	0.893	**32.22**	**0.920**
Average	35.23	0.941	34.44	0.940	33.54	0.897	34.13	0.917	33.68	0.934	34.94	0.937	34.70	0.935	**35.70**	**0.952**

Table 2. Quantitative comparison results of the proposed model with 13 other advanced models, including CNN and Transformer-based methods, on three datasets.

Methods	Rain200L		Rain200H		DID-Dataset		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RESCAN [21]	33.82	0.955	26.22	0.822	32.60	0.927	30.88	0.901
PreNet [29]	37.12	0.976	29.04	0.890	33.37	0.919	33.18	0.928
DRT [24]	38.81	0.983	28.67	0.880	33.88	0.928	33.79	0.930
CCN [28]	38.26	0.981	29.99	0.914	32.13	0.924	33.46	0.940
Restormer [44]	40.58	0.986	**30.83**	0.914	33.19	0.926	34.87	0.942
Uformer [38]	40.20	0.986	30.31	0.911	34.36	0.933	34.96	0.943
MPRNet [43]	39.82	0.986	29.94	0.900	34.50	0.937	34.75	0.941
SmartAssign [37]	38.41	0.981	27.71	0.854	33.11	0.915	33.08	0.917
SFNet [7]	39.50	0.982	29.75	0.901	34.51	0.938	34.59	0.940
NAFNet [2]	39.48	0.982	29.19	0.888	34.69	0.937	34.45	0.936
ELFformer [18]	38.85	0.980	28.93	0.885	33.54	0.936	33.77	0.934
ESDNet [31]	39.85	0.986	30.01	0.913	34.52	**0.939**	34.79	0.946
MSGNN [35]	39.09	0.987	29.63	**0.918**	33.11	0.927	33.94	0.944
ACL(Ours)	**40.74**	**0.988**	30.45	0.916	**34.81**	0.938	**35.33**	**0.947**

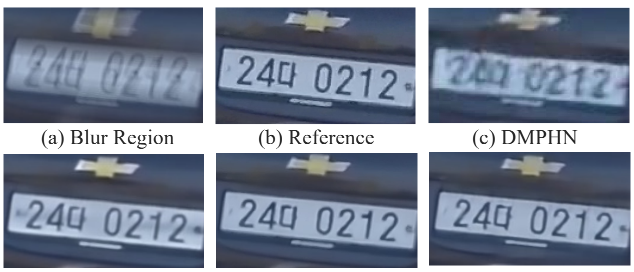


(a) Rainy Image (b) GT (c) MPRNet (d) MAXIM (e) Restormer (f) Ours

Figure 3. Visual comparison of ACL and other recent SOTA models on rainy image removal. The first two scenes contain slight rain streak degradation, while the last two scenes contain severe rain streak degradation.

Table 3. Quantitative results of our method compared to recent approaches in blurry image restoration.

Methods	GoPro			
	PSNR ↑	SSIM ↑	FLOPs(G) ↓	Param(M) ↓
MIMO [4]	32.68	0.959	617	16.1
DMPHN [47]	31.20	0.940	-	21.7
DBGAN [49]	31.10	0.942	759	11.6
MPRNet [43]	32.66	0.959	777	20.1
Restormer [44]	32.92	0.961	140	26.1
IRNeXt [6]	33.16	0.962	114	13.21
Stripformer [32]	33.08	0.962	170	20
SSAMAN [42]	33.53	0.965	165	18.3
LoFormer [26]	33.73	0.966	47	16.4
Ours	33.25	0.964	55	4.6



(a) Blur Region (b) Reference (c) DMPHN
(e) IRNeXt (f) Restormer (g) Ours

Figure 4. Visual results of ACL and four other advanced models on motion blurred image restoration.

be observed that our method outperforms the other compared methods, except for the powerful Transformer-based Restormer on the Rain200H. Additionally, Fig. 3 further illustrates the visual comparison results, where ACL demonstrates superior performance in restoring image details.

4.3. Single Image Deblurring Results

We conducted evaluations on motion blur image restoration using the GoPro dataset [27], which includes 2,103 training images and 1,111 testing images. The compared methods include the recently proposed 9 advanced methods. In Table 3, we present the comparative results of various metrics on this dataset. As shown, ACL achieves advanced performance while maintaining a low parameter count and low FLOPs. Additionally, Fig. 4 displays visual comparison results with other methods. We selected critical numerical information within the images, and it can be observed that ACL also exhibits good performance in restoring motion-blurred images.

4.4. Ablation Studies

To further understand the contributions of each module in the ACL model and other factors affecting model performance, we conducted a unified experiment on the Rain100L/H rain removal dataset. Specifically, we performed ablation studies on the two main modules of the model. Additionally, to verify that the linear attention capability can be restored using the Mamba structure, we conducted experiments by replacing LAMA with other structures to compare the results under different configurations.

Table 4. Comparison of ablation experiments between two modules on Rain100L.

Settings	PSNR ↑	SSIM ↑
Baseline	38.24	0.978
Baseline + LAMA	38.89	0.981
Baseline + LAMA + MDC	39.18	0.983

4.4.1. LAMA and MDC modules:

To validate the roles of the two main modules in ACL, namely LAMA and MDC, as well as their respective importance, we conducted ablation experiments. We removed the MDC module and replaced the core encoder/decoder modules with the structure shown in Fig. 1(b), where the Transformer block in the baseline model is implemented based on Linear Attention. We then gradually replaced the encoder-decoder modules with LAMA and added the MDC module to the baseline model, resulting in two different configurations, which were trained and tested separately. The experimental results are shown in Table 4. By comparing the “Baseline” and “Baseline+LAMA” configurations, it is evident that the Mamba structure, implemented with linear attention, achieves better results, proving the potential of the Mamba structure to activate linear attention. Besides, adding the MDC further enhances the model’s performance.

4.4.2. Different Mechanisms for Core Modules:

To validate the effectiveness of LAMA, we replaced the LA structure within LAMA with the original one-dimensional scanning SSM structure and the bidirectional scanning SSM structure for verification. Additionally, replacing LAMA with a standard LA structure resulted in a new configuration, referred to as the Baseline model. The results obtained from these various configurations are presented in Table 5. On one hand, the proposed method outperforms both the Baseline and the strategies employing unidirectional and multi-directional scanning for modeling. On the other hand, compared to the unidirectional Mamba model, the multi-directional scanning strategy demonstrates superior performance, indicating that unidirectional modeling is not optimal for image data. Furthermore, Fig. 5 provides visual examples corresponding to each configuration. It can be observed that while the SSM-based scanning methods are capable of removing rainy degradation elements, they result in significant loss of image details. In contrast, our method achieves superior visual outcomes.

Table 5. Quantitative results of different mechanisms used in the core module on rainy streak removal.

Settings	Rain100L		Rain100H	
	PSNR	SSIM	PSNR	SSIM
Baseline	38.24	0.978	31.02	0.913
Mamba (w. 1D Scan)	36.47	0.959	29.72	0.887
Mamba (w. 2D Scan)	38.07	0.977	30.93	0.907
Ours	39.18	0.983	32.22	0.920

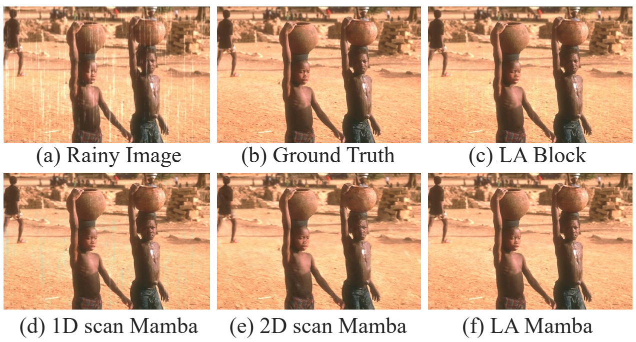


Figure 5. Visual results of the encoder/decoder blocks adopting different mechanisms.

Table 6. Evaluation results of different N_i configurations on the Rain100L dataset in ACL. ⋆ represents the combination we adopted.

(N_1,N_2,N_3)	PSNR↑	SSIM↑
(3,3,3)	38.95	0.978
(4,4,4)	39.03	0.980
(6,3,3)⋆	39.18	0.983
(6,6,6)	39.19	0.983

4.4.3. The Impact of the Number of Encoders/Decoders

As mentioned in the above, a crucial hyper-parameter in the model is the number of feature processing blocks at each encoding/decoding stage, denoted as N_i. To investigate the impact of N_i, we conducted a series of experiments on the deraining dataset using different combinations, with the results presented in Table 6. Additionally, Fig. 6 illustrates the output images generated by the models with various configurations. The numerical results indicate that increasing N_i contributes to performance improvement, but the effect is limited. Notably, a significant enhancement is observed when N_1 is increased, suggesting that learning high-resolution feature maps plays a critical role in improving the model's recovery performance.

4.4.4. The Dilation Rate of Convolution in MDC Module

We propose a local feature enhancement module, MDC, which includes two dilated convolution layers, as shown in Fig. 2 (d). We further investigate the impact of dilation rates on the module's performance. As shown in Table 7, the results indicate that compared to the non-dilated convolution setting ($d = 1$), the performance is inferior to the other two configurations. However, with increased dilation rates, performance improves to varying degrees, with the best overall performance observed at $d = 2$.

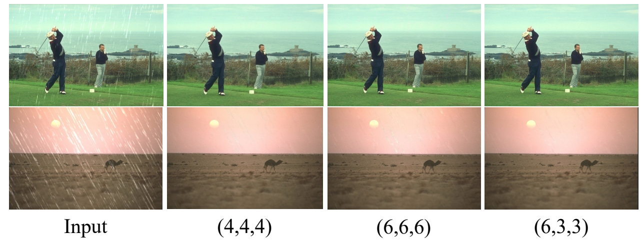


Figure 6. Visual Results on Various N_i

Table 7. The evaluation results of the MDC module's convolution operations with varying dilation rates on the Rain100L dataset. ⋆ represents the combination we adopted.

dilation	PSNR↑	SSIM↑
1	38.12	0.981
2⋆	39.18	0.983
3	39.10	0.981

4.5. Conclusion and Limitations

We introduced ACL, a novel image restoration model that addresses the limitations of traditional CNNs and Transformer-based approaches in handling global receptive fields and computational efficiency. By integrating linear attention into the Mamba structure, we developed the LAMA module, which enhances global feature dependencies with linear computational complexity. Additionally, the MDC module was designed to improve local detail restoration through multi-scale dilated convolutions. Our experiments confirmed that the ACL achieves promising performance in de-raining tasks, demonstrating its effectiveness in both quantitative metrics and visual quality. Furthermore, our work provides a new perspective on leveraging the Mamba structure in the IR domain. Nevertheless, our method also has some limitations. The ACL model does not have an advantage over CNN-based models when processing large-sized images. Moreover, in the image deblurring task, there is still a gap between ACL and SOTA Transformer models. In the future, the ACL model still has room for further optimization to adapt more IR tasks.

Acknowledgments

This work was supported by the National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U22B2051, No. 62302411), the Natural Science Foundation of Fujian Province of China (No.2021J06003), and China Postdoctoral Science Foundation (No. 2023M732948).

References

[1] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 3

[2] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022. 5, 6

[3] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. Learning a sparse transformer network for effective image deraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5896–5905, 2023. 3

[4] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 7

[5] Yuning Cui, Wenqi Ren, Xiaochun Cao, and Alois Knoll. Image restoration via frequency selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[6] Yuning Cui, Wenqi Ren, Sining Yang, Xiaochun Cao, and Alois Knoll. Irnext: Rethinking convolutional network design for image restoration. 2023. 1, 5, 6, 7

[7] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. Selective frequency network for image restoration. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[8] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 2, 3

[9] Enxuan Gu, Hongwei Ge, and Yong Guo. Code: An explicit content decoupling framework for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2920–2930, 2024. 2

[10] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2511–2520, 2019. 2

[11] Yubin Gu, Honghui Xu, Yueqian Quan, Wanjun Chen, and Jianwei Zheng. Orsi salient object detection via bidimensional attention and full-stage semantic guidance. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13, 2023. 1

[12] Yubin Gu, Siting Chen, Xiaoshuai Sun, Jiayi Ji, Yiyi Zhou, and Rongrong Ji. Optical remote sensing image salient object detection via bidirectional cross-attention and attention restoration. *Pattern Recognition*, page 111478, 2025. 1

[13] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024. 1, 2, 5, 6

[14] Tianyu Guo, Haowei Wang, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. Improving panoptic narrative grounding by harnessing semantic relationships and visual confirmation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1985–1993, 2024. 1

[15] Dongchen Han, Ziyi Wang, Zhuofan Xia, Yizeng Han, Yifan Pu, Chunjiang Ge, Jun Song, Shiji Song, Bo Zheng, and Gao Huang. Demystify mamba in vision: A linear attention perspective. *Advances in Neural Information Processing Systems*, 37:127181–127203, 2025. 2, 3

[16] Tao Huang, Xiaohuan Pei, Shan You, Fei Wang, Chen Qian, and Chang Xu. Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*, 2024. 3

[17] Jiayi Ji, Haowei Wang, Changli Wu, Yiwei Ma, Xiaoshuai Sun, and Rongrong Ji. Jm3d jm3d-llm: Elevating 3d representation with joint multi-modal cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2475–2492, 2025. 1

[18] Kui Jiang, Zhongyuan Wang, Chen Chen, Zheng Wang, Laizhong Cui, and Chia-Wen Lin. Magic elf: Image deraining meets association learning and transformer. *arXiv preprint arXiv:2207.10455*, 2022. 6

[19] Kui Jiang, Wenxuan Liu, Zheng Wang, Xian Zhong, Junjun Jiang, and Chia-Wen Lin. Dawn: Direction-aware attention wavelet network for image deraining. In *Proceedings of the 31st ACM international conference on multimedia*, pages 7065–7074, 2023. 6

[20] Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5886–5895, 2023. 2

[21] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018. 6

[22] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 2

[23] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF inter-*

national conference on computer vision, pages 1833–1844, 2021. 1, 2, 3

[24] Yuanchu Liang, Saeed Anwar, and Yang Liu. Drt: A lightweight single image deraining recursive transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 589–598, 2022. 3, 6

[25] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024. 3

[26] Xintian Mao, Jiansheng Wang, Xingran Xie, Qingli Li, and Yan Wang. Loformer: Local frequency transformer for image deblurring. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10382–10391, 2024. 7

[27] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 7

[28] Ruijie Quan, Xin Yu, Yuanzhi Liang, and Yi Yang. Removing raindrops and rain streaks in one go. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9147–9156, 2021. 6

[29] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3937–3946, 2019. 6

[30] Yuan Shi, Bin Xia, Xiaoyu Jin, Xing Wang, Tianyu Zhao, Xin Xia, Xuefeng Xiao, and Wenming Yang. Vmambair: Visual state space model for image restoration. *arXiv preprint arXiv:2403.11423*, 2024. 1, 2, 3

[31] Tianyu Song, Guiyue Jin, Pengpeng Li, Kui Jiang, Xiang Chen, and Jiyu Jin. Learning a spiking neural network for efficient image deraining. *IJCAI*, 2024. 6

[32] Fu-Jen Tsai, Yan-Tsung Peng, Yen-Yu Lin, Chung-Chi Tsai, and Chia-Wen Lin. Stripformer: Strip transformer for fast image deblurring. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 2, 7

[33] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5769–5780, 2022. 6

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2

[35] Cong Wang, Wei Wang, Chengjin Yu, and Jie Mu. Explore internal and external similarity for single image deraining with graph neural networks. *IJCAI*, 2024. 6

[36] Qiong Wang, Kui Jiang, Jinyi Lai, Zheng Wang, and Jianhui Zhang. Hpcnet: A hybrid progressive coupled network for image deraining. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2747–2752. IEEE, 2023. 1

[37] Yinglong Wang, Chao Ma, and Jianzhuang Liu. Smartassign: Learning a smart knowledge assignment strategy for deraining and desnowing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2023. 6

[38] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17683–17693, 2022. 2, 6

[39] Xinyu Xie, Yawen Cui, Tao Tan, Xubin Zheng, and Zitong Yu. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Visual Intelligence*, 2(1): 37, 2024. 3

[40] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1357–1366, 2017. 5

[41] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. Frequency and spatial dual guidance for image dehazing. In *European Conference on Computer Vision*, pages 181–198. Springer, 2022. 1

[42] Anas Zafar, Danyal Aftab, Rizwan Qureshi, Xinqi Fan, Pingjun Chen, Jia Wu, Hazrat Ali, Shah Nawaz, Sheheryar Khan, and Mubarak Shah. Single stage adaptive multi-attention network for image restoration. *IEEE TIP*, 2024. 7

[43] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14821–14831, 2021. 6, 7

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022. 1, 2, 5, 6, 7

[45] Wengyi Zhan, Mingbao Lin, Chia-Wen Lin, and Rongrong Ji. Anysr: Realizing image super-resolution as any-scale, any-resource. *IEEE Transactions on Image Processing*, 2024. 2

[46] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018. 5

[47] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5978–5986, 2019. 7

[48] Jinlu Zhang, Yiyi Zhou, Qiancheng Zheng, Xiaoxiong Du, Gen Luo, Jun Peng, Xiaoshuai Sun, and Rongrong Ji. Fast text-to-3D-aware face generation and manipulation via direct cross-modal mapping and geometric regularization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 60605–60625. PMLR, 2024. 1

[49] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2737–2746, 2020. 7

[50] Jianwei Zheng, Wei Li, Ni Xu, Junwei Zhu, Xiaoxu Lin, and Xiaoqin Zhang. Alias-free mamba neural operator. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3