Transformer-Inspired Convolutional Network for Image Restoration

Yuning Cui, Mingyu Liu, Wenqi Ren, Boxin Shi, and Alois Knoll

***Abstract*—Image restoration is a critical task in computer vision, as unwanted degradations significantly reduce image quality and adversely affect the robustness of downstream vision systems. Recently, Transformer-based models have demonstrated remarkable performance in image restoration tasks. However, their quadratic computational complexity with respect to input size limits their practical application. In this paper, we develop a computationally efficient convolutional image restoration model inspired by the design philosophy of Transformer blocks. Specifically, we emulate and unify the functionalities of self-attention and feed-forward layers within a single convolution-based module to achieve adaptive and efficient spatial feature learning. Additionally, this module incorporates multiple receptive fields at attention, module, and scale levels, enabling effective management of degradations of varying sizes. Furthermore, to mitigate frequency discrepancies between degraded and sharp image pairs, we widen and deepen the structure of a frequency learning mechanism, which implicitly refines frequency spectra through a simple yet powerful frequency decomposition and calibration approach. Integrating these designs into a U-shaped convolutional backbone results in our efficient and effective network, termed UIRNet, for image restoration. Extensive experiments across single-degradation, all-in-one, and composite degradation restoration tasks validate the effectiveness of our method. Notably, UIRNet achieves state-of-the-art performance on 11 benchmark datasets covering six single-degradation scenarios and performs favorably against leading all-in-one approaches under a five-degradation setting. Importantly, our method also demonstrates robust performance on two challenging composite degradation image restoration benchmarks, as well as in ultra-high-definition (UHD) applications.**

***Index Terms*—Image restoration, convolutional network, frequency refinement, multi-scale learning, Transformer-inspired architecture, all-in-one image restoration, composite degradation image restoration, ultra-high-definition image restoration**

I. Introduction

THE captured images often suffer from various degradations, such as haze, rain, and blur, due to the adverse conditions or physical limitations of cameras. Such degradations often hamper image visibility and degrade downstream vision systems' robustness. In this regard, image restoration can be beneficial as it aims to remove those undesired degradations and reconstruct a clean image from the corrupted inputs [1]. As such, image restoration plays an important role in many fields, such as remote sensing and autonomous vehicles. Many conventional algorithms have been developed for image restoration by exploring various assumptions and hand-crafted features to reduce the solution space. However, these methods are inapplicable to complicated real-world scenarios.

Recent convolutional neural networks (CNNs) have powerful learning capabilities for generalizable priors from large-scale datasets and have significantly advanced the performance of image restoration tasks [2], [3]. The convolution operator performs well at learning local connectivity, but it is restrained by the limited receptive field and is inadequate to remove large-scale degradations. In addition, the fixed convolution kernel is not flexible enough to adapt to the input that often involves non-uniform blurs [1].

In this regard, Transformer models have mitigated the aforementioned limitations, yielding promising performance across a diverse range of image restoration tasks [1], [4]. However, these Transformers have quadratic complexity to the input size despite their strong ability to model long-range pixel interactions effectively. This limitation restricts their broader applicability in image restoration tasks, which typically involve high-resolution inputs. In addition, the self-attention mechanism produces an attention matrix based on the local signals, causing sub-optimal information aggregation for degradation-aware large-scale degradation restoration. Moreover, the sum-to-one normalization method is utilized in self-attention for pixel integration and generates all-positive attention weights that are unable to filter detrimental pixels [5].

More recently, CNN-based architectures have experienced a resurgence in high-level vision tasks, driven by innovations involving large kernel sizes [6], multi-scale feature extraction [7], and Transformer-inspired training methodologies [8]. In line with this trend, this paper introduces a Transformer-inspired convolutional network for image restoration that emulates the core characteristics of Transformer blocks through computationally efficient convolutional operations. To this end, we imitate and unify the functionalities of self-attention and feed-forward networks [1] within a single convolution-based module. More specifically, for the attention part, we employ a lightweight branch to yield integration weights for pixels by taking into consideration global contexts. Then, we break the restriction of the sum-to-one operation by replacing Softmax with the hyperbolic tangent function [9]. In this way, we emphasize the informative pixels and suppress the harmful signals via the attention weights that have been projected to $(-1, 1)$. Additionally, we use the efficient convolution operator for integration, rather than the expensive dot-product of self-attention [5]. Finally, we embed multi-scale receptive fields into the attention mechanism by learning attention weights for different spatial sizes.

On the other hand, the gated-dconv feed-forward network [1] has demonstrated considerable effectiveness in image restoration tasks. In this work, we incorporate this gated mechanism into our Transformer-inspired module by employing

Corresponding author: Boxin Shi (shiboxin@pku.edu.cn)

Yuning Cui, Mingyu Liu, and Alois Knoll are with the School of Computation, Information and Technology, Technical University of Munich, Munich, Germany.

Wenqi Ren is with the School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen, Guangdong, China.

Boxin Shi is with the State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing, China.

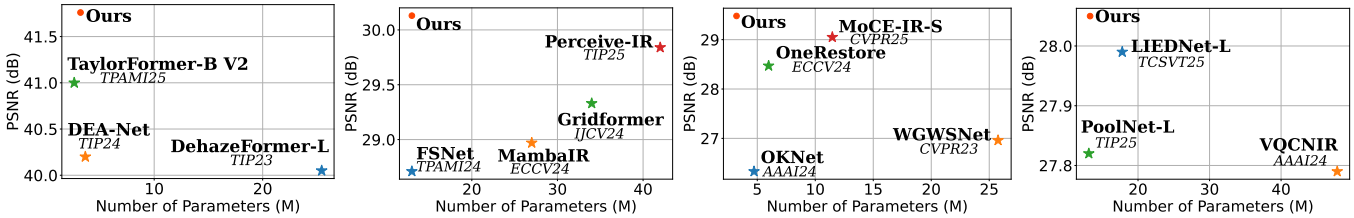


Fig. 1: Comparison of the number of parameters for single-degradation (a), all-in-one (b), and composite degradation (c,d) image restoration tasks. The performance of all-in-one methods is reported as the average over five tasks (see Table XVII).

a large-kernel depth-wise convolution in one branch, complemented by the previously described attention mechanism in the other branch. The outputs from these two branches are subsequently merged through element-wise multiplication. Furthermore, the kernel size is varied across different scales of the U-shaped backbone, facilitating multi-scale feature learning at the network's scale level.

In addition to enhancing feature learning in the spatial domain, we also develop a frequency learning module to refine spectral features by sufficiently leveraging the frequency gap between degraded/sharp image pairs. This module is established on a simple yet effective mechanism that achieves implicit frequency decoupling and recalibration using average pooling. In this work, we deepen and widen this mechanism to improve its ability for multi-scale and high-order frequency representation learning.

The main contributions are summarized as follows:

- We propose a Transformer-inspired module that integrates the functionalities of self-attention and feed-forward networks into a unified convolution-based component, enabling efficient spatial feature learning. The attention mechanism is implemented using convolutional operations, and the proposed module further incorporates multi-scale feature learning at the attention, module, and scale levels.
- We introduce a frequency learning module by widening and deepening a frequency decoupling and recalibration mechanism, thereby enabling multi-scale and high-order frequency representation learning.
- Extensive experiments conducted on three categories of image restoration tasks (single-degradation, all-in-one, and composite degradation)[1] validate the effectiveness and efficiency of our designs (see Figure 1). Moreover, our model can be effectively adapted to UHD image restoration tasks.

This paper is an extended version of [10]. The improvements include:

- In addition to the convolution-based spatial attention mechanism, we incorporate the functionality of a feed-forward network, resulting in a unified Transformer-inspired module. Moreover, the proposed module integrates multi-scale feature learning capabilities at the attention, module, and scale levels.
- We enhance the frequency learning mechanism by deepening and widening its structure to facilitate multi-scale and high-order frequency representation refinement.
- Beyond image motion deblurring, we extend our model to five more single-degradation image restoration tasks: image dehazing, deraining, denoising, defocus deblurring, and low-light enhancement. Furthermore, we evaluate our model's effectiveness in more challenging all-in-one, composite degradation, and UHD scenarios.

II. Related Work

A. Image Restoration

CNNs have demonstrated remarkable performance across various image restoration tasks [11]. Among these architectures, the encoder-decoder framework has emerged as a popular paradigm due to its effectiveness in learning hierarchical representations. Additionally, numerous advanced modules and mechanisms have been carefully designed and integrated into image restoration models to improve performance. For instance, to enable the network to focus on residual information learning, both image-level and feature-level skip connections have been widely adopted [2], [12]. Moreover, diverse channel and spatial attention mechanisms have been developed to selectively emphasize critical features [13], [14]. Extensive efforts have also been made to expand the receptive fields by deepening backbone structures and employing techniques such as dilated convolutions [15] and non-local blocks [16].

More recently, numerous Transformer-based models have emerged in image restoration [9], [17], inspired by their outstanding performance on high-level vision tasks. However, despite their effectiveness in capturing long-range dependencies, these models suffer from quadratic computational complexity associated with the self-attention mechanism. Several approaches have been proposed to mitigate this issue by limiting the spatial operation region [18] or altering the dimension of operations [1]. Nevertheless, these modifications still impose significant complexity constraints with respect to specific dimensions such as window size or channel count. In this study, we propose a Transformer-inspired module

[1] **Single-degradation** or **task-specific** image restoration methods are designed to effectively address a specific type of degradation once trained. As a result, separate models must be trained for different degradation types. In contrast, **all-in-one** algorithms can handle multiple degradation types and severity levels after a single training process, with each image in the dataset affected by only one type of degradation. Meanwhile, **composite degradation** or **mixed-degradation** image restoration methods aim to remove multiple degradations that occur simultaneously within a single image.

that integrates the functionalities of self-attention and feed-forward layers into a unified convolution-based unit, where the attention mechanism is realized through convolutional operations to ensure computational efficiency. Additionally, we introduce carefully designed multi-scale learning mechanisms into both the spatial and frequency modules to further enhance the model's representational learning capability.

B. All-in-One Image Restoration

Single-degradation image restoration has attracted considerable research attention in recent years. However, deploying separate models for different degradations is impractical and resource-intensive, particularly for resource-constrained platforms. Consequently, all-in-one solutions have recently gained significant interest due to their capability to address multiple degradations using a single unified model trained only once [19]. A common practice in such methods is to initially extract degradation-specific information from input images, which subsequently guides the restoration process. For instance, AirNet [20], one of the pioneering approaches, contrastively learns degradation representations from corrupted images and leverages these representations to reconstruct clean images. However, the effectiveness of contrastive learning largely depends on carefully constructed positive and negative pairs. PromptIR [21] encodes degradation representations into learnable prompts, which are integrated into decoders for image reconstruction, resulting in increased parameter overhead. More recently, text-based prompts generated from large models have been utilized to achieve controllable [22] reconstruction. Nevertheless, the involvement of large models significantly increases computational complexity and overhead. Despite the absence of highly specialized components, our proposed model achieves competitive performance against state-of-the-art all-in-one frameworks, highlighting its strong representational learning capability. Furthermore, we conduct extensive experiments across three major categories of image restoration tasks, a comprehensive evaluation rarely explored in previous studies.

III. Proposed Method

This section first presents the overall pipeline of the Transformer-inspired convolutional network, termed UIRNet. Then, we delineate the core components: Transformer-inspired module (TIM) and frequency learning module (FLM). Finally, the loss functions used are introduced.

A. Overall Pipeline

A schematic of the proposed UIRNet is shown in Figure 2. Following recent schemes [2], [12], UIRNet applies the popular encoder-decoder solution for image restoration and contains three resolution scales to learn hierarchical representations. In the case of a degraded image $\mathbf{I}_1 \in \mathbb{R}^{3\times H\times W}$, a 3×3 convolutional layer is used to extract shallow features of size $C \times H \times W$, where C denotes the number of channels and $H \times W$ represents spatial locations. Next, the shallow features are fed into ResBlocks, which consist of n normal

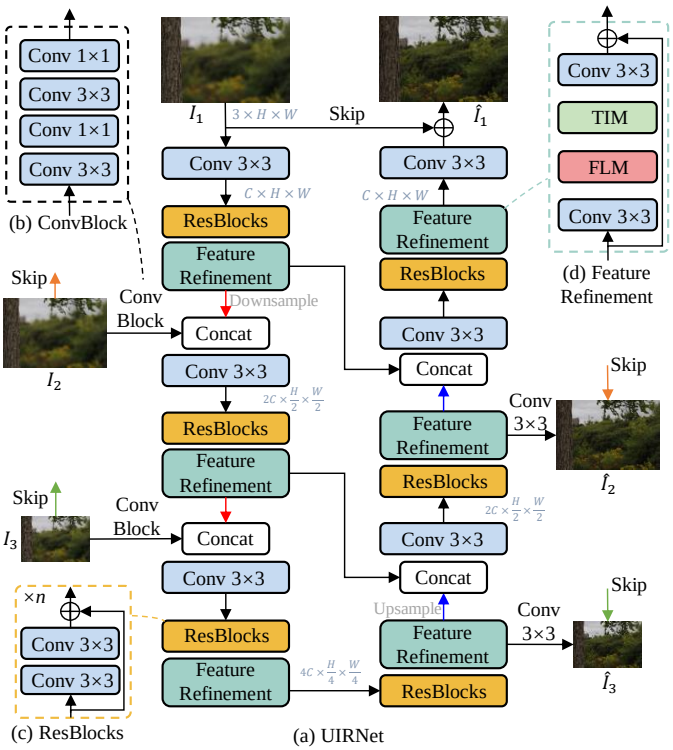


Fig. 2: UIRNet adopts a three-scale U-shaped architecture and employs multi-input/output strategies [2]. Our proposed components are incorporated into a regular residual block.

residual blocks (Figure 2 (c)), and the feature refinement module (Figure 2 (d)). Then, the multi-input strategy [2], [13] is adopted, where the low-resolution input image $\mathbf{I}_2$ is injected into the main path via concatenation after being processed by ConvBlock (Figure 2 (b)). The channel dimension is adjusted by a 3×3 convolution. Subsequently, the resulting features pass through another two combinations of ResBlocks and feature refinement modules to obtain the deepest features with the shape of $4C \times \frac{H}{4} \times \frac{W}{4}$. This process involves $\mathbf{I}_3$. Then, the deepest features are processed by the decoder with three scales to yield a sharp image. The decoder features are concatenated with the encoder features through feature-level skip connections to facilitate the recovery process. The model outputs low-resolution predicted images via 3×3 convolutions after the first two feature refinement modules in the decoder network. For each pair of same-size input/output images, the image-level skip connection is utilized following [2]. The proposed two components, TIM and FLM, are deployed between two 3×3 convolutional layers, as shown in Figure 2 (d).

B. Transformer-Inspired Module (TIM)

TIM is designed to emulate the operation of Transformer blocks, aiming to enhance spatial representational learning while maintaining high efficiency. To achieve this, the canonical self-attention is first revisited to reveal its defects.

1) Attention Mechanism: The formulation of multi-head self-attention (MHSA) [23] can be expressed as:

$$\mathrm{MHSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathrm{head}_1, ..., \mathrm{head}_h]W^O, \qquad (1)$$

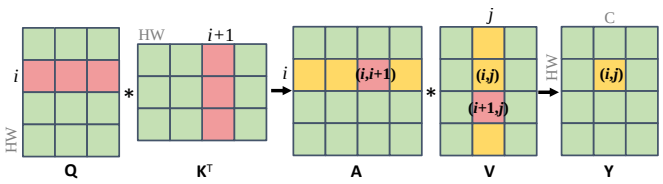


Fig. 3: Calculation process of self-attention [23].

$$\text{head}_i = \text{Attention}(\mathbf{Q}W_i^Q, \mathbf{K}W_i^K, \mathbf{V}W_i^V), \tag{2}$$

where $[\cdot,\cdot]$ denotes the concatenation operation. W_i^Q,W_i^K, and W_i^V are parameter matrices to project $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ into different feature spaces. The scaled dot-product attention is the core component that is given by:

$$\begin{aligned}\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \mathbf{AV} \\ &= \text{Softmax}(\mathbf{QK}^\top)\mathbf{V}.\end{aligned} \tag{3}$$

Here, we omit the normalization factor for simplicity. Given an input of size $\mathbb{R}^{HW\times C}$, the sizes of the attention map $\mathbf{A}$ and outcome are $HW \times HW$ and $HW \times C$, respectively. For convenience, we have assumed here that $\mathbf{Q},\mathbf{K},\mathbf{V} \in \mathbb{R}^{HW\times C}$ have the same number of channels as the input. Next, we demonstrate that the attention weights of self-attention are generated only from local interactions, as illustrated in Figure 3. Specifically, for a single resulting pixel (i,j) generated by Eq. 3, its response is attained by:

$$\mathbf{Y}_{i,j} = \sum_{p=1}^{HW} \mathbf{A}_{i,p}\mathbf{V}_{p,j}. \tag{4}$$

The contribution of $\mathbf{V}_{i+1,j}$ to the calculation of $\mathbf{Y}_{i,j}$ is determined by $\mathbf{A}_{i,i+1}$, which is yielded by:

$$\mathbf{A}_{i,i+1} = \sum_{q=1}^{C} \mathbf{Q}_{i,q}\mathbf{K}^\top_{q,i+1}. \tag{5}$$

As demonstrated in the analyses above, we utilize only the source information from these two spatial locations to compute the contribution of pixel $(i+1,j)$ to (i,j). Although self-attention enables global value aggregation, its affinity scores are mainly generated from pairwise query-key similarities and are not explicitly conditioned on image-level degradation context. This may make the aggregation weights less tailored to degradation-aware high-resolution image restoration, where effective filtering often depends on both global degradation statistics and local structural reliability.

To ameliorate this issue, the attention weights in TIM are produced based on global information. More concretely, as shown in Figure 4, TIM leverages global average pooling (GAP) to generate the global feature, and then a convolution layer is used to obtain the initial attention weights. We impose the hyperbolic tangent activation function on the initial weights to produce factors falling into $(-1,1)$, instead of using Softmax in the standard self-attention [24]. As a consequence, the harmful pixels can be suppressed appropriately [25]. Formally, given $X \in \mathbb{R}^{C\times H\times W}$, our attention weights are produced by:

$$\mathbf{W} = \text{Tanh}(f_{1\times1}(\text{GAP}(\mathbf{X}))), \tag{6}$$

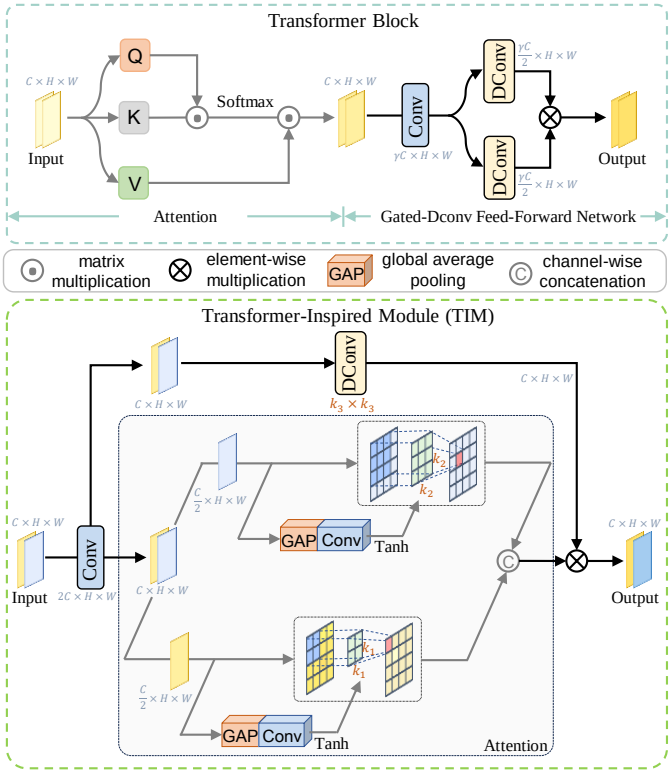


Fig. 4: Illustrations of the Transformer Block [1] and our proposed Transformer-inspired module (TIM). Conv and DConv indicate convolution and depth-wise convolution, respectively.

where $f_{1\times1}$ is a point-wise convolution. Here, we can either generate a similar attention matrix as self-attention and perform matrix multiplication with inputs, or produce the kernel values of the convolution operator for aggregation. Since the former still has quadratic complexity, we adopt the more efficient integration method based on convolution [26]. Here, $\mathbf{W} \in \mathbb{R}^{1\times k^2}$, where $k\times k$ is the kernel size. Next, the refined feature can be generated after reshaping $\mathbf{W}$ to $\mathbb{R}^{k\times k}$:

$$\mathbf{Y}_{c,h,w} = \sum_{i=0}^{k-1}\sum_{j=0}^{k-1} \mathbf{W}_{i,j}\mathbf{X}_{c,h-\lfloor\frac{k}{2}\rfloor+i,w-\lfloor\frac{k}{2}\rfloor+j} + \mathbf{X}_{c,h,w}. \tag{7}$$

The above process can be concluded as $\mathbf{Y} = \mathcal{A}_k(\mathbf{X})$. As shown in the attention mechanism in Figure 4, the input is split into two components, *i.e.*, $\mathbf{X}_1, \mathbf{X}_2 = \mathcal{S}(\mathbf{X})$, where $\mathcal{S}$ denotes the splitting operation. Spatial refinement is then applied to each component using different kernel sizes k, yielding multi-scale receptive fields.

Moreover, inspired by the group-query attention [27] that shares single key and value heads for each group of query heads, our spatial refinement is carried out in groups in each split part to strike a balance between the diversity of attention weights and efficiency (see Figure 5). Specifically, the split features ($\mathbf{X}_1$ and $\mathbf{X}_2$) are divided into several groups along the channel dimension, and distinct attention weights are produced for each group. The weights are shared among both spatial and channel dimensions in each group to improve efficiency. To summarize, the entire process of the attention mechanism in

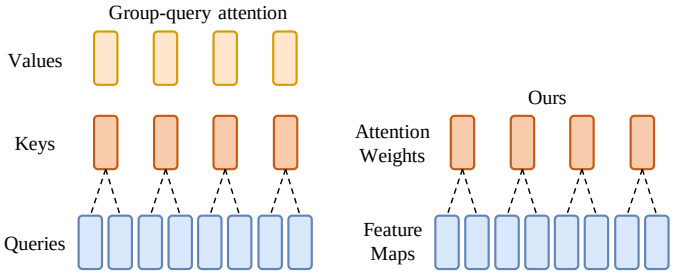


Fig. 5: Illustrations of group-query attention [27] in Transformer models and our attention design.

TIM is given by:

$$\hat{\mathbf{X}} = [[\mathcal{A}_3^1(\mathbf{X}_{1,1}), ..., \mathcal{A}_3^g(\mathbf{X}_{1,g})], [\mathcal{A}_5^1(\mathbf{X}_{2,1}), ..., \mathcal{A}_5^g(\mathbf{X}_{2,g})]], \tag{8}$$

$$\mathbf{X}_{1,1}, ..., \mathbf{X}_{1,g} = \mathcal{S}(\mathbf{X}_1), \quad \mathbf{X}_{2,1}, ..., \mathbf{X}_{2,g} = \mathcal{S}(\mathbf{X}_2), \tag{9}$$

where the superscript means using separate attention weights for different groups, and g is the number of groups specified in each component, which is set to 4 in our model.

2) Integrating Feed-Forward Network Design: In addition to mimicking self-attention, we integrate the design philosophy of the feed-forward network (FFN) into our module to further enhance representation learning. To achieve this, we choose a well-established gated-dconv feed-forward network [1] in image restoration for emulation. As illustrated in the top right corner of Figure 4, this FFN first uses a convolution layer to expand the feature channels by a factor of γ. 3×3 depth-wise convolutions are used in parallel branches obtained by splitting the expanded features. The output is generated by merging two branches via element-wise multiplication.

Inspired by this mechanism, we first apply a 1×1 convolution to double the number of feature channels. After splitting, one branch undergoes a depth-wise convolution with a kernel size of $k_3 \times k_3$, while the other branch is processed through our proposed attention mechanism. Finally, the two branches are fused via element-wise multiplication.

Moreover, in our case, k_3 increases with the scale number and is defined as $k_3 = 7 + 2 \times scale$, where $scale \in \{0, 1, 2\}$ denotes the scale index of the model[2]. Consequently, TIM incorporates three levels of multi-scale learning: attention level, module level, and scale level.

Discussion: We note that TIM is built upon several widely used primitive operations, such as dynamic convolution, depth-wise convolution, and gated modulation. Therefore, the novelty of TIM does not lie in inventing each primitive operator independently. Instead, TIM aims to reorganize these operations into a restoration-oriented functional structure. Existing Transformer-inspired CNNs mainly use convolutional operations as efficient replacements for self-attention or as stand-alone modulation units [5], [6]. In contrast, TIM embeds the attention-like dynamic aggregation branch into an FFN-style multiplicative formulation. Specifically, one branch performs content-adaptive regional aggregation using globally conditioned dynamic convolution, while the other branch performs local spatial modulation using depth-wise convolution. The two branches are coupled through element-wise multiplication, allowing regional interaction and local modulation to be jointly performed within a single compact module. Moreover, TIM incorporates receptive-field diversity at the attention, module, and scale levels, which is particularly suitable for image restoration degradations with varying spatial extents.

[2]The first scale corresponds to the highest feature resolution.

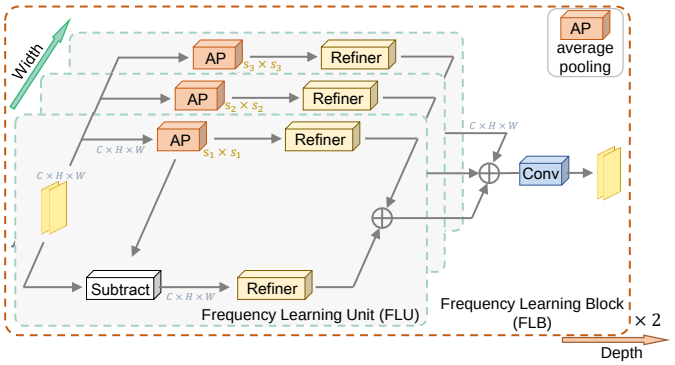


Fig. 6: Architectural details of our Frequency Learning Module (FLM). $s \times s$ denotes the kernel size of pooling.

C. Frequency Learning Module (FLM)

Apart from enhancing spatial feature learning, we further improve frequency learning by effectively leveraging the discrepancies between corrupted and clear image pairs. Most frequency-based approaches first leverage the fast Fourier transform or Wavelet transform to yield spectral features, and then resort to convolution layers to process features [15], [28], [29]. In this study, we instead adopt a simple yet effective frequency decomposition and refinement method as our basic unit to process different frequency components individually [10]. The architectural details of FLM are illustrated in Figure 6. We begin by detailing the basic frequency learning unit (FLU) and subsequently introduce our expansion strategy to facilitate multi-scale and high-order frequency learning.

Concretely, given $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, FLU uses the average pooling technique to extract the low-frequency feature, and the corresponding high-frequency feature is yielded by subtracting the resulting low frequencies from the input. This process can be formally expressed as:

$$\mathbf{X}_l = \mathrm{AP}_s(\mathbf{X}), \quad \mathbf{X}_h = \mathbf{X} - \mathbf{X}_l, \tag{10}$$

where $\mathbf{X}_l$, $\mathbf{X}_h$ are low- and high-frequency components, respectively; AP_s denotes average pooling with the kernel size of $s \times s$. Subsequently, refiners are utilized to enhance the extracted features. Specifically, the features are multiplied by attention parameters, which are directly optimized through backpropagation, to recalibrate different frequency components effectively. The refined features are then added, as expressed below:

$$\mathbf{Y} = W_l \mathbf{X}_l + W_h \mathbf{X}_h, \tag{11}$$

where W_l and W_h are channel-wise attention weights. We conclude the above calculation process as $\mathbf{Y} = \mathcal{F}_s(\mathbf{X})$.

TABLE I: Image dehazing comparisons on both the synthetic (SOTS [30]) and real-world (Dense-Haze [31]) datasets.

	SOTS-indoor		SOTS-outdoor		Dense-Haze		Overhead	
Methods	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	Params/M↓	FLOPs/G↓
DeHamer [4]	36.63	0.988	35.18	0.986	16.62	0.560	132.45	60.3
MAXIM-2S [13]	38.11	0.991	34.19	0.985	-	-	14.1	108
HCD [32]	38.31	<u>0.995</u>	34.02	<u>0.994</u>	16.41	0.566	5.58	104.03
DehazeFormer-L [17]	40.05	**0.996**	-	-	-	-	25.44	279.7
TaylorFormer-B V1 [33]	40.71	0.992	37.42	0.989	16.66	0.560	2.68	38.5
DEA-Net [34]	40.20	0.993	36.03	0.989	-	-	3.653	32.23
TaylorFormer-B V2 [35]	<u>41.00</u>	0.993	<u>37.81</u>	0.991	<u>16.95</u>	<u>0.621</u>	2.63	37.7
Ours	**41.76**	**0.996**	**38.18**	**0.995**	**17.02**	**0.630**	3.21	31.5

Long- and short-range contexts play an important role in image restoration [1], [9]. Therefore, we widen FLU to form a frequency learning block (FLB) by employing pooling of different square kernel sizes in multiple units for frequency refinement, followed by a convolution operation to further enhance the added features. The whole computational process of FLB can be summarized as:

$$\begin{aligned}\hat{\mathbf{X}} &= \mathrm{FLB}(\mathbf{X}) \\ &= f_{3\times3}(\mathcal{F}_{GAP}(\mathbf{X}) + \sum_{i=1}^{4}\mathcal{F}_{2i+1}(\mathbf{X})) + \mathbf{X},\end{aligned} \tag{12}$$

where $\mathcal{F}_{GAP}$ means using GAP for frequency separation in FLU. As shown in Eq. 12, global and several local square pooling techniques are utilized to provide global-to-local multi-scale receptive fields for frequency refinement. Finally, we empirically deepen FLB to establish our FLM to achieve high-order frequency learning.

Discussion: The proposed FLM extends a single-level decomposition-and-calibration unit to a multi-scale and high-order refinement module. The widened structure decomposes features using multiple pooling scales, providing global-to-local frequency components. The deepened structure performs iterative refinement, where later frequency learning blocks operate on recalibrated features and further reduce residual frequency discrepancies. Therefore, FLM is not a simple repetition of the previous frequency unit, but a progressive frequency refinement mechanism. The spectral visualizations in Figure 10 show that successive FLBs progressively align the intermediate feature spectra with the ground-truth spectrum.

D. Loss Function

The l_1 loss in both spatial and frequency domains [2] is leveraged to facilitate dual-domain learning:

$$\begin{aligned}\mathcal{L}_s &= \sum_{i=1}^{3}\frac{1}{E_i}\|\hat{\mathbf{I}}_i - \mathbf{G}_i\|_1, \\ \mathcal{L}_f &= \sum_{i=1}^{3}\frac{1}{E_i}\|\mathrm{FFT}(\hat{\mathbf{I}}_i) - \mathrm{FFT}(\mathbf{G}_i)\|_1,\end{aligned} \tag{13}$$

where FFT is fast Fourier transform; $\hat{\mathbf{I}}_i$, $i \in \{1, 2, 3\}$ denotes the multi-scale predicted images as shown in Figure 2; $\mathbf{G}_i$ is the ground-truth image; and E represents the total pixel elements for normalization. The final loss function is given by combining the above two terms: $\mathcal{L} = \mathcal{L}_s + 0.1\mathcal{L}_f$.

IV. Experimental Results

We conduct extensive experiments on single-degradation, all-in-one, composite degradation, and UHD image restoration tasks to validate the effectiveness and generality of the proposed image restoration architecture.

- Single-degradation: the model is trained separately for each image restoration task, covering 11 datasets across image dehazing, deraining, denoising, motion deblurring, defocus deblurring, and low-light image enhancement;
- All-in-one: a single unified model is trained on a mixed dataset constructed from multiple datasets, where each image contains only one type of degradation;
- Composite degradation: the model aims to restore clean images from inputs in which each image is simultaneously corrupted by multiple types of degradations;
- UHD: the model is extended to UHD applications, supporting image restoration at a resolution of 3840×2160.

Such a broad evaluation is intended to demonstrate the generality and robustness of the proposed architectural design, rather than optimizing for a single benchmark. Details on the datasets, the specific training configuration, and additional visual comparisons are provided in the Supplementary Material. In the tables, the best results are highlighted in **bold**, while the second-best results are <u>underlined</u>.

A. Single-Degradation Image Restoration Results

1) Image Dehazing: We evaluate our model on both synthetic (SOTS [30]) and real-world datasets (Dense-Haze [31]). As shown in Table I, our method exhibits the best quality performance in three datasets. Specifically, our method outperforms DeHamer [4] on the SOTS-indoor dataset by 5.13 dB PSNR with only 2.4% parameters. Moreover, our model obtains significant performance improvements over the recent TaylorFormer [33], [35] and DEA-Net [34] models with similar computation overhead. In addition, our model is evaluated in challenging real-world scenarios using the Dense-Haze [31] dataset. Table I shows that our model performs favorably against state-of-the-art algorithms on Dense-Haze. Specifically, our approach outperforms TaylorFormer-B V2 [35] by 0.07 dB PSNR while consuming lower complexity.

Given the crucial role of image dehazing in remote sensing, we evaluate our model on the SateHaze1k [48] dataset to assess its performance in this domain. The quantitative results are shown in Table IV. Our model demonstrates strong performance across all three haze levels. Notably, it outperforms

TABLE II: Single-image defocus deblurring comparisons on the DPDD [36] dataset.

Methods	Indoor Scenes				Outdoor Scenes				Combined			
	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓	PSNR↑	SSIM↑	MAE↓	LPIPS↓
LaKDNet [37]			-				-		26.15	0.810	-	**0.155**
Restormer [1]	28.87	0.882	0.025	**0.145**	23.24	0.743	0.050	**0.209**	25.98	0.811	0.038	0.178
FocalNet [12]	29.10	0.876	0.024	0.173	23.41	0.743	**0.049**	0.246	26.18	0.808	0.037	0.210
MambaIR [38]	28.89	0.879	0.026	0.171	23.36	0.738	0.051	0.243	26.11	0.809	0.039	0.202
OKNet [39]	28.99	0.877	0.024	0.169	23.51	0.751	**0.049**	0.241	26.18	0.812	0.037	0.206
RDDM [40]			-				-		25.97	0.811	0.037	0.166
Ours	**29.21**	**0.883**	**0.023**	0.155	**23.58**	**0.755**	**0.049**	0.212	**26.32**	**0.817**	**0.036**	0.185

TABLE III: Low-light image enhancement comparisons on the LOL-v2-syn [41] dataset.

Methods	Uformer [9]	Restormer [1]	MIRNet [42]	SNR-Net [43]	Retinexformer [44]	MambaLLIE [45]	FourierMamba [46]	CIDNet [47]	**Ours**
PSNR↑	19.66	21.41	21.94	24.14	25.67	25.87	24.75	25.71	**25.95**
SSIM↑	0.871	0.830	0.876	0.928	0.930	0.940	0.945	0.942	**0.954**

TABLE IV: Image dehazing results on the remote sensing dataset, SateHaze1k [48], which includes thin, moderate, and thick haze levels.

Methods	Thick		Moderate		Thin	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
UMWTransformer [49]	20.07	0.825	26.65	0.946	24.29	0.919
Trinity-Net [50]	20.97	0.823	23.35	0.895	21.55	0.884
FocalNet [12]	21.69	0.847	25.99	0.947	24.16	0.916
FMambaIR [51]	22.65	0.850	25.83	0.939	**24.58**	0.912
Ours	**23.03**	**0.957**	**26.70**	**0.978**	24.40	**0.977**

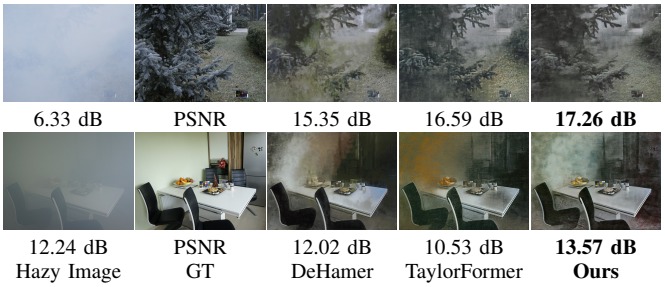


Fig. 7: Dehazing results on the Dense-Haze [31] dataset.

TABLE V: Image deraining results on the AGAN-Data [52] dataset for raindrop removal.

Methods	IDT [52]	MAXIM-2S [53]	AWRCP [13]	FPro [54]	AST-B [55]	**Ours** [56]
PSNR↑	31.87	31.87	31.93	31.96	32.32	**32.74**
SSIM↑	0.931	0.935	0.931	0.937	0.935	**0.943**

the specialized algorithm [50] by a significant margin of 2.06 dB in thick haze scenarios.

Figure 7 shows that the images produced by our model are much visually closer to ground truth, suggesting the robustness of our model in addressing real-world degradations.

2) Image Defocus Deblurring: The DPDD [36] dataset is used to compare our model with other state-of-the-art algorithms for single-image defocus deblurring. Table II shows that our model outperforms other algorithms on most metrics. Specifically, our method obtains a performance improvement

TABLE VI: Image motion deblurring results on GoPro [57].

Method	GoPro		Overhead	
	PSNR↑	SSIM↑	Params/M↓	FLOPs/G↓
MIMO-UNet+ [2]	32.45	0.957	16.1	154.41
MPRNet [14]	32.66	0.959	20.1	777.01
MAXIM [13]	32.86	0.961	22.2	169.5
Restormer [1]	32.92	0.961	26.13	141
DDANet [10]	33.07	0.962	16.18	153.51
PromptRestorer [58]	33.06	0.962	24.4	186
Stripformer [18]	33.08	0.962	20	170.46
TaylorFormer-XL V2 [35]	33.24	**0.963**	16.26	141.9
Ours	**33.28**	**0.963**	13.3	126

of 0.34 dB in PSNR over the strong Transformer model Restormer [1] in three categories. In addition, our model achieves a performance gain of 0.14 dB over the recent CNN-based method OKNet [39] in the combined category while using fewer parameters (see Figure 1c). The comparisons demonstrate the effectiveness of the proposed method.

3) Image Deraining: We evaluate the proposed model on the widely adopted AGAN-Data [52] dataset for raindrop removal. The quantitative results are presented in Table V. As shown, our method achieves a performance gain of 0.42 dB over the recent Transformer-based AST-B [56] while maintaining significantly lower computational complexity, as illustrated in Figure 1 (b). Additionally, our model outperforms the prompt-based FPro [55] by 0.78 dB in PSNR while reducing complexity by 61%.

4) Low-Light Image Enhancement: For this task, we evaluate the proposed method on the widely adopted LOL-v2-syn [41] dataset. The quantitative results are summarized in Table III, showing that our model achieves competitive performance compared to state-of-the-art algorithms. Notably, our approach outperforms the recent Mamba-based FourierMamba [46] and Transformer-based Retinexformer [44], with PSNR improvements of 0.20 dB and 0.28 dB, respectively. In addition, our model surpasses CIDNet [47], which performs enhancement in the Horizontal/Vertical-Intensity (HVI) space.

5) Image Motion Deblurring: The proposed model is evaluated on the commonly adopted GoPro [57] dataset for motion

Fig. 8: Visualization of intermediate feature maps in TIM.

TABLE VII: Comparison of inference speed across four image restoration tasks using an NVIDIA Tesla V100 GPU. Our model achieves better performance with faster speed.

Tasks	Datasets	Methods	PSNR↑	Time/s↓
Dehazing	SOTS-indoor	TaylorFormer-B V1	40.71	0.843
		Ours	**41.76**	0.087
Low Light	LOL-v2-syn	Retinexformer	25.67	0.054
		Ours	**25.95**	0.049
Denoising	BSD68	Restormer	26.62	0.277
		Ours	**26.65**	0.092
Deblurring	GoPro	Stripformer	33.08	1.054
		Ours	**33.28**	0.587

TABLE VIII: Peak inference memory footprint comparison at different input resolutions on an NVIDIA RTX 4090 GPU. "-" indicates out-of-memory.

Resolution	1200×1200	1600×1600	2000×2000
TaylorFormer-B V2	10352.68 MB	18385.72 MB	-
Ours	6017.34 MB	10673.72 MB	16665.03 MB

deblurring. Table VI shows that our model performs favorably against previous methods with fewer parameters and lower complexity. Specifically, our model achieves a performance improvement of 0.36 dB PSNR over Restormer [1], while utilizing only 51% of its parameters. Furthermore, compared to the recent Transformer-based TaylorFormer-XL V2 [35], our model achieves a 0.04 dB gain in PSNR while reducing parameters by 18% and computational complexity by 11%.

6) Inference Time Comparison: We compare the inference time of our model with well-established algorithms across four tasks. Table VII demonstrates that our model outperforms Transformer-based models by achieving a higher score while maintaining superior inference speed. Notably, on the SOTS-indoor [30] dataset, our model achieves nearly 10× faster inference speed compared to TaylorFormer-B [33].

7) Memory Footprint Comparison: We further compare the memory footprint of the proposed model with that of a recent state-of-the-art method, TaylorFormer-B V2 [35], under different input resolutions. As shown in Table VIII, our model consistently requires less peak inference memory as the input resolution increases. Notably, TaylorFormer-B V2 runs out of memory at a resolution of 2000×2000, whereas our model can still perform inference successfully.

TABLE IX: Ablation study of each module.

	Baseline	TIM	FLM	PSNR	Params	FLOPs
(a)	✓			31.33	1.48M	15.44G
(b)	✓	✓		36.40	1.64M	16.77G
(c)	✓		✓	36.85	1.50M	15.64G
(e)	✓	✓	✓	37.32	1.66M	16.98G

TABLE X: Ablation study on the integration of FFN.

Methods	w/o FFN (a)	FFN+fixed kernel (b)	FFN+multi-scale kernel (c)
PSNR↑	35.19	36.22	36.40

8) Ablation Study: Ablation studies are conducted by training dehazing models on the RESIDE-indoor [30] dataset so as to verify the effectiveness of our modules and investigate different options. All configurations remain unchanged, except for n in Figure 2 (c), which is set to 0. The baseline model is obtained by directly removing TIM and FLM from our model.

Effects of each component. Table IX presents a breakdown ablation study in which the baseline model receives 31.33 dB on the SOTS-indoor [30] dataset. TIM and FLM achieve performance gains of 5.07 dB and 5.52 dB, respectively. Here, TIM introduces only 0.16M parameters and 1.33G FLOPs, while FLM introduces only 0.02M parameters. The model, equipped with both modules, obtains a further performance gain of 5.99 dB compared to the baseline. The results suggest the effectiveness of our modules.

In addition, we visualize the intermediate feature maps of TIM. Figure 8 shows that the features input to TIM primarily capture edges, but the representations are relatively unclear. In contrast, the features processed by TIM not only strengthen edge signals but also improve the overall consistency of the feature maps, indicating that TIM effectively leverages both local and large-scale information to refine the representations.

Multi-scale learning in TIM. Table X demonstrates that removing the FFN design from TIM results in a PSNR of 35.19 dB, leading to a performance drop of 1.21 dB. The scale-level multi-scale learning mechanism leads to a performance gain of 0.18 dB PSNR over the fixed version, which employs 7×7 depth-wise convolutions in all scales.

We further analyze the impact of varying kernel sizes in the attention mechanism. As illustrated in Table XI, the performance improves with the increase in kernel size. In addition, Table XIe and Table XIf explore the benefits of using multi-scale receptive fields. The input features are split along the channel dimension to reduce complexity. The 3×3 kernel is imposed on half of the channels, while the other half is divided equally for the other kernels. Taking Table XIf as an example, the 3×3 kernel is used on half channels of the feature, and $5 \times 5/7 \times 7$ kernel is used on a quarter of channels. In this case, the number of groups is set to 4, 2, and 2 for the 3×3, 5×5, and 7×7 kernels, respectively. The use of 3×3 and 5×5 helps to obtain 35.19 dB PSNR, which is higher than that of using a separate one. Moreover, this choice reduces the training time compared with that of 5×5 kernel. Finally, kernel sizes of 3×3 and 5×5 are chosen as our configuration

TABLE XI: Ablation study of kernel sizes in TIM. Models are trained on an NVIDIA Tesla A100 GPU. The total number of groups is set as 8.

Kernel Size	3×3 (a)	5×5 (b)	7×7 (c)	9×9 (d)	3×3/5×5 (e)	3×3/5×5/7×7 (f)
PSNR↑	34.59	34.97	35.58	35.74	35.19	35.24
Training Time/h↓	11.72	17.30	24.72	36.51	15.27	18.09

TABLE XII: Ablation study of the number of groups in TIM.

Groups	2 (a)	4 (b)	8 (Table XIa) (c)	16 (d)	32 (e)
PSNR↑	34.13	34.17	34.59	34.46	34.22

TABLE XIII: Ablation study on the effect of receptive field size on the attention weights in TIM.

Receptive field	PSNR	Params	FLOPs
$\frac{H}{2} \times \frac{W}{2}$	37.18	1.75M	17.02G
Global	37.32	1.66M	16.98G

TABLE XIV: Ablation study of activation functions in TIM.

Activation Function	None (a)	Relu (b)	Sigmoid (c)	Softmax (d)	Tanh (Table XIId) (e)
PSNR↑	34.41	34.33	34.23	34.34	34.46

TABLE XV: Ablation study of multi-scale receptive fields in FLB.

	Global	3×3	5×5	7×7	9×9	PSNR↑
(a)						31.33
(b)	✓					34.00
(c)		✓				33.13
(d)	✓	✓				34.67
(e)	✓	✓	✓			35.47
(f)	✓	✓	✓	✓		35.70
(g)	✓	✓	✓	✓	✓	35.86

to strike a better balance between performance and efficiency.

Different numbers of groups in TIM. The diversity of attention weights in TIM is enhanced by increasing the number of groups, and the results are reported in Table XII. When the number of groups is increased from 2 to 8, the performance improves from 34.13 to 34.59. As the number of groups increases further, the accuracy declines. Therefore, a total of 8 groups is selected for our final model.

Effect of receptive field size on the attention weights in TIM. The TIM generates attention weights based on a global receptive field, in contrast to the self-attention mechanism, which relies on a local receptive field to compute attention weights. To validate this design choice, we conduct an ablation experiment by reducing the receptive field in TIM during attention weight computation. Specifically, we first apply an adaptive average pooling operator to produce a contextual feature map of size $C \times 2 \times 2$, which is then reshaped to $\mathbb{R}^{4C\times1\times1}$ and passed through a 1×1 convolution to generate the attention weights. As shown in Table XIII, this variant achieves 37.18 dB PSNR, which is 0.14 dB lower than our original design, while also incurring higher computational overhead. These results validate the effectiveness of leveraging a global receptive field in TIM for attention weight generation.

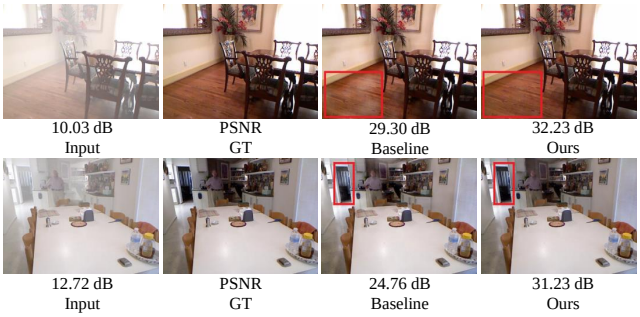


Fig. 9: Image dehazing comparisons on SOTS-indoor [30] between models with and without using our method.

TABLE XVI: Ablation study on deepening FLB.

Methods	FLB	2*FLB (=FLM)	3*FLB
PSNR↑	35.86	36.85	36.95

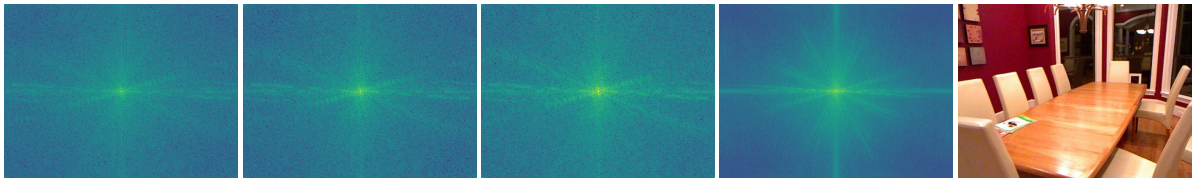
Different activation functions. The influences of different activation functions in TIM are also studied, and their results are presented in Table XIV. Interestingly, the model obtains the second-best performance without using any activation function (Table XIVa). When Tanh is used to project weights to (-1, 1), the model exhibits the best performance. Therefore, the Tanh function is used as the default function.

Ablation studies on FLM. We first investigate the effects of widening FLU in FLB. Table XV shows that, compared to the baseline model, the GAP variant (Table XVb) produces a performance gain of 2.67 dB PSNR while the local 3×3 variant (Table XVc) obtains a gain of 1.8 dB PSNR. The model (Table XVd), using both global and 3×3 local frequency refinement, further boosts the performance to 34.67 dB. The performance improves from 34.67 dB to 35.86 dB when more local operations with different kernel sizes are used. Note that input features are not split into groups as TIM due to the high efficiency of implementation for average pooling. The version of Table XVg is selected in our model.

In Figure 9, the produced images by models Table XV (a) and (g) are compared. When equipped with our method, the model becomes more effective in removing haze blurs, such as blurs on the floor and the door.

We further examine the impact of expanding FLB along the depth dimension. Table XVI indicates that increasing the number of FLB enhances performance but nearly saturates when two FLB are stacked. Therefore, we incorporate two FLB in each FLM in our model.

In addition, we visualize the spectra of intermediate features in the FLM in Figure 10. As observed, the two FLBs progressively recover the frequency components, bringing them closer to those of the ground truth, particularly in the high-frequency regions. Furthermore, we compute the spectra of the restored results for the second example in Figure 7. As shown in Figure 11, compared with competing Transformer-based approaches, the spectrum produced by our model more

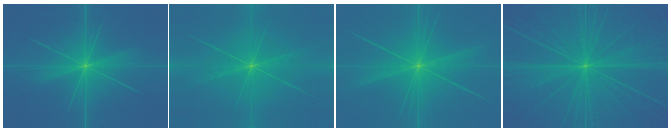


Before FLM | After the first FLB | After the second FLB | GT (Spectrum) | GT

Fig. 10: Spectra of intermediate features in FLM (consisting of two FLBs), illustrating that the proposed FLB progressively aligns the frequency components with those of the ground truth.

TABLE XVII: Results for the all-in-one setting, where a single model is trained on a mixed dataset comprising five types of degradations. "+" indicates the number of parameters introduced by more modalities.

	Dehazing on SOTS		Deraining on Rain100L		Denoising on BSD68		Deblurring on GoPro		Low-Light on LOL		Average		
Methods	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	Params↓
Restormer [1]	24.09	0.927	34.81	0.962	31.49	0.884	27.22	0.829	20.41	0.806	27.60	0.881	26M
FSNet [59]	25.53	0.943	36.07	0.968	31.33	0.883	28.32	0.869	22.29	0.829	28.71	0.898	13M
IDR [60]	25.24	0.943	35.63	0.965	**31.60**	0.887	27.87	0.846	21.34	0.826	28.34	0.893	15M
PromptIR [21]	26.54	0.949	36.37	0.970	31.47	0.886	28.71	0.881	22.68	0.832	29.15	0.904	33M
MambaIR [38]	25.81	0.944	36.55	0.971	31.41	0.884	28.61	0.875	22.49	0.832	28.97	0.901	27M
Gridformer [61]	26.79	0.951	36.61	0.971	31.45	0.885	29.22	0.884	22.59	0.831	29.33	0.904	34M
InstructIR [22]	27.10	0.956	36.84	0.973	31.40	0.887	29.40	**0.886**	**23.00**	0.836	29.55	0.907	16M
Perceive-IR [19]	28.19	0.964	37.25	0.977	31.44	0.887	**29.46**	**0.886**	22.88	0.833	29.84	0.909	42+86M
AdaPrompt-IR [62]	30.33	**0.977**	**37.27**	**0.978**	31.26	0.888	28.33	0.872	22.59	0.836	29.96	0.910	34M
Ours	**30.76**	**0.977**	36.97	**0.978**	31.25	**0.889**	28.75	0.869	22.91	**0.847**	**30.13**	**0.912**	13M



DeHamer | TaylorFormer | **Ours** | GT

Fig. 11: Comparison of the frequency spectra corresponding to the second example in Fig. 7.

closely matches that of the ground truth, demonstrating its superior capability in restoring fine image details.

B. All-in-One Image Restoration Results

We assess the effectiveness of our model in an all-in-one image restoration setting [20]–[22], [63], where a single model is trained on a compound dataset collected from five tasks, including dehazing, deraining, denoising, motion deblurring, and low-light image enhancement. Once trained, the model can effectively address multiple types of degradations.

The results are reported in Table XVII. Our model performs well against state-of-the-art algorithms on most datasets while requiring fewer parameters. Particularly on the SOTS [30] dataset for image dehazing, our method achieves a significant performance gain of 0.43 dB PSNR compared to the second-best AdaPrompt-IR [62]. Moreover, averaged over five tasks, our model achieves the highest performance, outperforming Perceive-IR by 0.29 dB in PSNR while retaining strong parameter efficiency. Notably, despite not explicitly incorporating degradation priors, our model achieves superior overall performance compared to the recent AdaPrompt-IR [62], which constructs a degradation representation codebook to distinguish and capture degradation components. These results underscore the effectiveness of our model in addressing multiple degradations using a single model instance.

Discussion. The goal of the all-in-one setting is to address multiple types of degradations using a single unified model, *i.e.,* achieving strong overall performance across all evaluated categories. Our model demonstrates significantly superior performance on certain tasks, *e.g.,* dehazing, while underperforming competing methods on others, such as deblurring. This behavior is common in multi-task learning, where different models tend to excel at different tasks due to inherent discrepancies among tasks and architectural biases. Such differences make it challenging for a single model to achieve optimal performance across all tasks simultaneously.

C. Composite Degradation Image Restoration Results

To evaluate the robustness of the proposed method in composite degradation scenarios, we conduct experiments under the three-degradation setting on CDD-11 [64] and the two-degradation setting on LOL-Blur [68].

1) Three-Degradation Setting: We train and evaluate our model on the CDD-11 [64] dataset, which consists of images affected by up to three types of degradation simultaneously. This dataset includes a total of 11 degradation categories, generated through various combinations of four primary degradation types: low-light, rain, haze, and snow. The quantitative comparisons are presented in Table XVIII. Our model significantly outperforms the task-specific [1], [39] and all-in-one competing approaches. Compared to the recent composite degradation restoration framework, OneRestore [64], our model obtains a remarkable performance gain of 1.01 dB PSNR while saving $\sim 50\%$ parameters. In addition, compared

TABLE XVIII: Composite degradation image restoration on CDD-11 [64], with performance reported in PSNR and SSIM. "-" indicates that the results are unavailable.

Method	Params	Low (L)		Haze (H)		Rain (R)		Snow (S)		L+H		L+R		L+S		H+R		H+S		L+H+R		L+H+S		Average	
OKNet [39]	5M	25.64	.813	28.27	.983	30.98	.947	32.35	.963	24.08	.808	24.54	.773	24.20	.766	25.83	.934	27.14	.949	23.13	.766	23.48	.764	26.33	.861
AirNet [20]	9M	24.83	.778	24.21	.951	26.55	.891	26.79	.919	23.23	.779	22.82	.710	23.29	.723	22.21	.868	23.29	.901	21.80	.708	22.24	.725	23.75	.814
PromptIR [21]	36M	26.32	.805	26.10	.969	31.56	.946	31.53	.960	24.49	.789	25.05	.771	24.51	.761	24.54	.924	23.70	.925	23.74	.752	23.33	.747	25.90	.850
OneRestore [64]	6M	26.48	.826	32.52	.990	33.40	.964	34.31	.973	25.79	.822	25.58	.799	25.19	.789	29.99	.957	30.21	.964	24.78	.788	24.90	.791	28.47	.878
MoCE-IR-S [65]	11M	27.26	.824	32.66	.990	34.31	**.970**	**35.91**	**.980**	26.24	.817	26.25	.800	26.04	.793	29.93	**.964**	30.19	**.970**	25.41	.789	25.39	.790	29.05	.881
CPL [66]	35M	**27.35**	-	32.79	-	**34.60**	-	35.35	-	26.18	-	26.32	-	26.10	-	30.40	-	29.95	-	25.40	-	25.35	-	29.07	-
PoolNet-S [67]	4M	27.08	.829	33.16	.991	33.65	.965	34.88	.973	26.42	.828	26.15	.803	25.73	.794	29.99	.962	30.38	.965	25.22	.796	25.20	.791	28.90	.882
Ours	3M	27.34	**.832**	**33.74**	**.992**	34.16	.967	35.56	.976	**26.74**	**.831**	**26.43**	**.807**	**26.26**	**.801**	**31.24**	**.964**	**31.11**	**.970**	**25.82**	**.801**	**25.87**	**.800**	**29.48**	**.885**

13.08 dB Input | 23.82 dB OKNet | 21.85 dB AirNet | 24.16 dB OneRestore | **24.97 dB Ours** | PSNR GT

Fig. 12: Visual comparisons on the CDD-11 [64] dataset.

8.36 dB Input | 24.45 dB LEDNet | 25.90 dB VQCNIR | **28.91 dB Ours** | PSNR GT

Fig. 13: Visual results on LOL-Blur [68]. Our model exhibits superior capability in recovering details from blurred images, such as text on billboards.

TABLE XIX: Comparison between the sequential operation and our one-step method on composite degradations (haze and snow) using CDD-11 [64].

Method	Desnow→Dehaze	Dehaze→Desnow	Ours
PSNR	25.54	27.87	28.55
SSIM	0.919	0.958	0.954

with CPL [66], which enhances degradation-aware representations via prompt-based modules, our method achieves notable performance improvements, particularly on composite degradation combinations, indicating the strong representational capacity of our model. Furthermore, our model outperforms the frequency-based PoolNet-S [67] by an average of 0.58 dB in PSNR while using fewer parameters.

The visual comparisons in Figure 12 illustrate that our model is more effective in recovering structural details and removing degradation under composite conditions.

We further compare our method with sequential alternatives that remove individual degradations using separate models. Table XIX presents the results on images with a combination of haze and snow. In the sequential approach, each task-specific model is trained solely on a single degradation type, whereas our model is trained directly on images containing composite degradations. As shown in Table XIX, applying the dehazing model to desnowed images yields only 25.54 dB in PSNR. The reversed variant, "Dehaze→Desnow", significantly improves performance, indicating that the order of task-specific models is critical when addressing composite degradations. Our model outperforms both sequential variants in PSNR by handling composite degradations in a one-step manner, which also reduces training and inference time. This result suggests that direct learning on images with multiple degradation types better captures the interactions between degradations and is more likely to achieve higher performance.

TABLE XX: Quantitative results on LOL-Blur [68] for low-light image deblurring.

Methods	PSNR↑	SSIM↑	Params/M↓
MIMOUNet [2]	22.41	0.835	6.80
LLFlow [69]	24.48	0.846	-
LEDNet [68]	25.74	0.850	7.41
Restormer [1]	26.72	0.902	26.13
VQCNIR [70]	27.79	0.875	47.85
LIEDNet-L [71]	27.99	0.897	17.79
PoolNet-L [67]	27.82	0.913	13.12
Ours	**28.05**	**0.958**	13.3

2) Two-Degradation Setting: The quantitative results on the LOL-Blur [68] dataset for low-light image deblurring are summarized in Table XX. Our model outperforms state-of-the-art methods, including the general image restoration approach [1], dedicated deblurring method [2], and low-light enhancement techniques [69]. Notably, compared to the specialized algorithm VQCNIR [70], our model achieves a substantial improvement of 0.26 dB in PSNR and 0.083 in SSIM. The visual comparisons in Figure 13 further highlight

TABLE XXI: Quantitative comparisons on four UHD image restoration tasks.

Task (Data)	Methods	PSNR↑	SSIM↑
Dehazing (UHD-Haze [72])	UHDFormer [72]	22.59	0.943
	UHDDIP [74]	24.70	0.952
	ERR [73]	25.12	0.950
	Ours	**25.38**	**0.959**
Deraining (4K-Rain13k [76])	UDR-S2Former [75]	33.36	0.946
	UDR-Mixer [76]	34.30	0.951
	ERR [73]	34.48	0.952
	Ours	**34.63**	**0.954**
Desnowing (UHD-Snow [74])	SFNet [77]	23.63	0.845
	UHDformer [72]	36.61	0.988
	UHDDIP [74]	41.56	0.990
	Ours	**42.18**	**0.992**
Deblurring (UHD-Blur [72])	FFTformer [78]	25.41	0.725
	UHDFormer [72]	28.82	0.844
	ERR [73]	29.72	0.861
	Ours	**30.36**	**0.874**

Input/9.19 dB ERR/18.77 dB **Ours/22.52 dB** GT/PSNR

Fig. 14: Visual results on UHD-Haze [72] for image dehazing.

Hazy Image **Ours** GT

Fig. 15: Dehazing results on the real-world dataset [31].

the advantages of our approach, showcasing its ability to recover finer structural details from low-light blurred inputs.

D. UHD Image Restoration Results

In addition to general image restoration tasks, we further extend our model (tiny version) to UHD scenarios. Specifically, the model is trained separately on four UHD image restoration tasks: dehazing, deraining, desnowing, and deblurring. The quantitative results are presented in Table XXI. As shown, our model achieves the best performance across all four tasks in terms of both PSNR and SSIM. Notably, on the UHD-Blur [72] dataset for motion deblurring, our method outperforms the specialized ERR [73] algorithm by 0.64 dB in PSNR and 0.013 in SSIM.

Figure 14 shows that our model more effectively recovers details and colors from an extremely challenging hazy input.

V. Conclusion and Outlook

In this study, we introduce UIRNet, an efficient and effective convolutional network inspired by Transformer architectures. Specifically, our Transformer-inspired module consolidates the core components of Transformer blocks into a single module, where the self-attention mechanism is implemented using computationally efficient convolutional operators, and the feed-forward network design is leveraged to enhance representation learning. To further improve performance, we incorporate three levels of multi-scale learning. Additionally, we extend both the width and depth of a simple yet effective frequency learning unit to facilitate multi-scale and high-order frequency refinement. Extensive experiments on single-degradation, all-in-one, composite degradation, and UHD image restoration tasks validate the effectiveness of our model.

We provide additional visual results on the real-world dehazing dataset [31] in Fig. 15. A noticeable gap remains compared to the ground truth in terms of reconstruction details and color fidelity. This limitation primarily stems from the scarcity of real-world training data, which is inherently difficult to collect. Promising directions to address this issue include leveraging domain adaptation techniques to better exploit synthetic data, incorporating the strong generalization capabilities of large-scale models, and collecting more real-world data for training.

On the other hand, although UIRNet achieves promising results, its current architecture still has limitations. For example, FLM relies on average-pooling-based frequency decomposition, which is efficient but relatively coarse and not explicitly directional or content-adaptive. Therefore, it may be less effective for complex real-world frequency discrepancies, such as directional artifacts. Future work will explore directional frequency decomposition to improve robustness in real-world scenarios.

VI. Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant No. 62136001, No. U24B20175, No. 62322216), Beijing Natural Science Foundation (Grant No. L233024), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Grant No. Z241100003524012), and Scientific Research Innovation Capability Support Project for Young Faculty (Grand No.SRICSPYF-ZY2025012).

References

[1] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[2] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 4641–4650.

[3] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[4] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5812–5820.

[5] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Revitalizing convolutional network for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9423–9438, 2024.

[6] Q. Hou, C.-Z. Lu, M.-M. Cheng, and J. Feng, "Conv2former: A simple transformer-style convnet for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[7] Y. Chen, X. Yuan, J. Wang, R. Wu, X. Li, Q. Hou, and M.-M. Cheng, "Yolo-ms: rethinking multi-scale representation learning for real-time object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[8] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[9] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 683–17 693.

[10] Y. Cui, Y. Tao, W. Ren, and A. Knoll, "Dual-domain attention for image deblurring." in *Association for the Advancement of Artificial Intelligence*, 2023.

[11] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2034–2042.

[12] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Focal network for image restoration," in *ICCV*, 2023.

[13] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5769–5780.

[14] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 821–14 831.

[15] W. Zou, M. Jiang, Y. Zhang, L. Chen, Z. Lu, and Y. Wu, "Sdwnet: A straight dilated network with wavelet transformation for image deblurring," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2021, pp. 1895–1904.

[16] J. Li, W. Tan, and B. Yan, "Perceptual variousness motion deblurring with light global context refinement," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 4116–4125.

[17] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Transactions on Image Processing*, vol. 32, pp. 1927–1941, 2023.

[18] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *Proceedings of the European Conference on Computer Vision*, 2022.

[19] X. Zhang, J. Ma, G. Wang, Q. Zhang, H. Zhang, and L. Zhang, "Perceive-ir: Learning to perceive degradation better for all-in-one image restoration," *IEEE Transactions on Image Processing*, 2025.

[20] B. Li, X. Liu, P. Hu, Z. Wu, J. Lv, and X. Peng, "All-in-one image restoration for unknown corruption," in *CVPR*, 2022.

[21] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. Shahbaz Khan, "Promptir: Prompting for all-in-one image restoration," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[22] M. V. Conde, G. Geigle, and R. Timofte, "Instructir: High-quality image restoration following human instructions," in *European Conference on Computer Vision*, 2024, pp. 1–21.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] Y. Gu, Y. Meng, X. Sun, J. Ji, W. Ruan, and R. Ji, "Mixed degradation image restoration via local dynamic optimization and conditional embedding," *arXiv preprint arXiv:2411.16217*, 2024.

[25] K. Purohit, M. Suin, A. N. Rajagopalan, and V. N. Boddeti, "Spatially-adaptive image restoration using distortion-guided networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2021, pp. 2309–2319.

[26] Q. Han, Z. Fan, Q. Dai, L. Sun, M.-M. Cheng, J. Liu, and J. Wang, "On the connection between local attention and dynamic depth-wise convolution," *arXiv preprint arXiv:2106.04263*, 2021.

[27] J. Ainslie, J. Lee-Thorp, M. De Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai, "Gqa: Training generalized multi-query transformer models from multi-head checkpoints," *arXiv preprint arXiv:2305.13245*, 2023.

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

[29] C. Li, C. Guo, M. Zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, "Embedding fourier for ultra-high-definition low-light image enhancement." in *International Conference on Learning Representations*, 2023.

[30] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[31] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images," in *IEEE International Conference on Image Processing*, 2019, pp. 1014–1018.

[32] T. Wang, G. Tao, W. Lu, K. Zhang, W. Luo, X. Zhang, and T. Lu, "Restoring vision in hazy weather with hierarchical contrastive learning," *arXiv preprint arXiv:2212.11473*, 2022.

[33] Y. Qiu, K. Zhang, C. Wang, W. Luo, H. Li, and Z. Jin, "Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 12 802–12 813.

[34] Z. Chen, Z. He, and Z.-M. Lu, "Dea-net: Single image dehazing based on detail-enhanced convolution and content-guided attention," *IEEE Transactions on Image Processing*, 2024.

[35] Z. Jin, Y. Qiu, K. Zhang, H. Li, and W. Luo, "Mb-taylorformer v2: improved multi-branch linear transformer expanded by taylor formula for image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[36] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 111–126.

[37] L. Ruan, M. Bemana, H.-p. Seidel, K. Myszkowski, and B. Chen, "Revisiting image deblurring with an efficient convnet," *arXiv preprint arXiv:2302.02234*, 2023.

[38] H. Guo, J. Li, T. Dai, Z. Ouyang, X. Ren, and S.-T. Xia, "Mambair: A simple baseline for image restoration with state-space model," in *ECCV*, 2024.

[39] Y. Cui, W. Ren, and A. Knoll, "Omni-kernel network for image restoration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 2, 2024, pp. 1426–1434.

[40] H. Feng, H. Zhou, T. Ye, S. Chen, and L. Zhu, "Residual diffusion deblurring model for single image defocus deblurring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 3, 2025, pp. 2960–2968.

[41] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 2072–2086, 2021.

[42] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Learning enriched features for real image restoration and enhancement," in *Proceedings of the European conference on computer vision*, 2020, pp. 492–511.

[43] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2022, pp. 17 714–17 724.

[44] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-stage retinex-based transformer for low-light image enhancement," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023, pp. 12 504–12 513.

[45] J. Weng, Z. Yan, Y. Tai, J. Qian, J. Yang, and J. Li, "Mamballie: Implicit retinex-aware low light enhancement with global-then-local state space," *Advances in Neural Information Processing Systems*, 2024.

[46] D. Li, Y. Liu, X. Fu, J. Huang, S. Xu, Q. Zhu, and Z.-J. Zha, "Fouriermamba: Fourier learning integration with state space models for image deraining," in *Forty-second International Conference on Machine Learning*, 2025.

[47] Q. Yan, Y. Feng, C. Zhang, G. Pang, K. Shi, P. Wu, W. Dong, J. Sun, and Y. Zhang, "Hvi: A new color space for low-light image enhancement," in *CVPR*, 2025.

[48] B. Huang, L. Zhi, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using sar image prior based on conditional generative adversarial networks," in *WACV*, 2020.

[49] A. Kulkarni, S. S. Phutke, and S. Murala, "Unified transformer network for multi-weather image restoration," in *ECCV*, 2022.

[50] K. Chi, Y. Yuan, and Q. Wang, "Trinity-net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond," *TGRS*, 2023.

[51] X. Luan, H. Fan, Q. Wang, N. Yang, S. Liu, X. Li, and Y. Tang, "Fmambair: A hybrid state space model and frequency domain for image restoration," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[52] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *CVPR*, 2018.
[53] J. Xiao, X. Fu, A. Liu, F. Wu, and Z.-J. Zha, "Image de-raining transformer," *PAMI*, 2022.
[54] T. Ye, S. Chen, J. Bai, J. Shi, C. Xue, J. Jiang, J. Yin, E. Chen, and Y. Liu, "Adverse weather removal with codebook priors," in *ICCV*, 2023.
[55] S. Zhou, J. Pan, J. Shi, D. Chen, L. Qu, and J. Yang, "Seeing the unseen: A frequency prompt guided transformer for image restoration," in *ECCV*, 2024.
[56] S. Zhou, D. Chen, J. Pan, J. Shi, and J. Yang, "Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration," in *CVPR*, 2024.
[57] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
[58] C. Wang, J. Pan, W. Wang, J. Dong, M. Wang, Y. Ju, and J. Chen, "Promptrestorer: A prompting image restoration method with degradation perception," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8898–8912, 2023.
[59] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Image restoration via frequency selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[60] J. Zhang, J. Huang, M. Yao, Z. Yang, H. Yu, M. Zhou, and F. Zhao, "Ingredient-oriented multi-degradation learning for image restoration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
[61] T. Wang, K. Zhang, Z. Shao, W. Luo, B. Stenger, T. Lu, T.-K. Kim, W. Liu, and H. Li, "Gridformer: Residual dense transformer with grid structure for image restoration in adverse weather conditions," *International Journal of Computer Vision*, pp. 1–23, 2024.
[62] W. Sun, Q. Wang, Y. Wang, Z. Hou, Q. Yan, and Y. Zhang, "Adaprompt-ir: Adaptive learning to perceive degradation semantic and prompting for all-in-one image restoration," *Pattern Recognition*, vol. 169, p. 111875, 2026.
[63] Y. Cui, S. W. Zamir, S. Khan, A. Knoll, M. Shah, and F. S. Khan, "Adair: Adaptive all-in-one image restoration via frequency mining and modulation," in *The Thirteenth International Conference on Learning Representations*, 2025.
[64] Y. Guo, Y. Gao, Y. Lu, H. Zhu, R. W. Liu, and S. He, "Onerestore: A universal restoration framework for composite degradation," in *European Conference on Computer Vision*, 2024, pp. 255–272.
[65] E. Zamfir, Z. Wu, N. Mehta, Y. Tan, D. P. Paudel, Y. Zhang, and R. Timofte, "Complexity experts are task-discriminative learners for any image restoration," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 753–12 763.
[66] G. Wu, J. Jiang, K. Jiang, X. Liu, and L. Nie, "Beyond degradation redundancy: Contrastive prompt learning for all-in-one image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 4, pp. 4005–4022, 2026.
[67] Y. Cui, W. Ren, and A. Knoll, "Exploring the potential of pooling techniques for universal image restoration," *IEEE Transactions on Image Processing*, 2025.
[68] S. Zhou, C. Li, and C. Change Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *ECCV*, 2022.
[69] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot, "Low-light image enhancement with normalizing flow," in *AAAI*, 2022.
[70] W. Zou, H. Gao, T. Ye, L. Chen, W. Yang, S. Huang, H. Chen, and S. Chen, "Vqcnir: Clearer night image restoration with vector-quantized codebook," in *AAAI*, 2024.
[71] M. Liu, Y. Cui, W. Ren, J. Zhou, and A. C. Knoll, "Liednet: A lightweight network for low-light enhancement and deblurring," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
[72] C. Wang, J. Pan, W. Wang, G. Fu, S. Liang, M. Wang, X.-M. Wu, and J. Liu, "Correlation matching transformation transformers for uhd image restoration," in *AAAI*, 2024.
[73] C. Zhao, Z. Chen, Y. Xu, E. Gu, J. Li, Z. Yi, Q. Wang, J. Yang, and Y. Tai, "From zero to detail: Deconstructing ultra-high-definition image restoration from progressive spectral perspective," in *CVPR*, 2025.
[74] L. Wang, C. Wang, J. Pan, X. Liu, W. Zhou, X. Sun, W. Wang, and Z. Su, "Ultra-high-definition image restoration: New benchmarks and a dual interaction prior-driven solution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 36, no. 2, pp. 2052–2068, 2026.
[75] S. Chen, T. Ye, J. Bai, E. Chen, J. Shi, and L. Zhu, "Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks," in *ICCV*, 2023.
[76] H. Chen, X. Chen, C. Wu, Z. Zheng, J. Pan, and X. Fu, "Towards ultra-high-definition image deraining: A benchmark and an efficient method," *IEEE Transactions on Multimedia*, pp. 1–13, 2026.
[77] Y. Cui, Y. Tao, Z. Bing, W. Ren, X. Gao, X. Cao, K. Huang, and A. Knoll, "Selective frequency network for image restoration." in *International Conference on Learning Representations*, 2023.
[78] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *CVPR*, 2023.

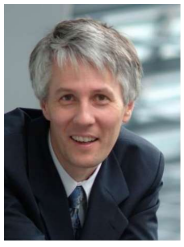
VII. Biography Section

Yuning Cui (Graduate Student Member, IEEE) received the B.Eng. degree from Central South University, China, in 2016 and the M.Eng. degree from National University of Defense Technology, China, in 2018. He is currently working towards a Ph.D. degree at the Chair of Robotics, Artificial Intelligence and Real-time Systems within the School of Computation, Information and Technology at the Technical University of Munich. His research interest lies in image restoration.

Mingyu Liu (Student Member, IEEE) is currently a PhD candidate in the Chair of Robotics, Artificial Intelligence and Real-time Systems at the Technical University of Munich (TUM), Germany. He received his dual master's degree in Electrical and Computer Engineering from TUM and Electronics and Communication Engineering from Tongji University, China, respectively. His research interests include computer vision in autonomous driving, deep learning, and artificial intelligence.

Wenqi Ren (Senior Member, IEEE) received the Ph.D. degree from Tianjin University, China, in 2017. From 2015 to 2016, he worked with Prof. Ming-Husan Yang as a Joint-Training Ph.D. Student with the Electrical Engineering and Computer Science Department, University of California at Merced. He is currently a Professor with the School of Cyber Science and Technology, Sun Yat-sen University, China. His research interests include image processing and related high-level vision problems.

Boxin Shi (Senior Member, IEEE) received the BE degree from the Beijing University of Posts and Telecommunications, in 2007, the ME degree from Peking University, in 2010, and the PhD degree from the University of Tokyo, in 2013. He is currently a Boya Young fellow associate professor (with tenure) and research professor with Peking University, where he leads the Camera Intelligence Lab. Before joining PKU, he did research with MIT Media Lab, Singapore University of Technology and Design, Nanyang Technological University, National Institute of Advanced Industrial Science and Technology, from 2013 to 2017. His papers were awarded as Best Paper, Runners-Up at CVPR 2024, ICCP 2015, and selected as Best Paper candidate at ICCV 2015. He is an associate editor of *IEEE Transactions on Pattern Analysis and Machine Intelligence/International Journal of Computer Vision* and an area chair of CVPR/ICCV/ECCV.



Alois Knoll (Fellow, IEEE) received his Ph.D. (*summa cum laude*) in Computer Science from Technical University of Berlin, Germany, in 1988. Since 2001, he has been a professor at the Department of Informatics, Technical University of Munich (TUM), Germany. His research interests include adaptive systems, multimedia information retrieval, model-driven development of embedded systems with applications to automotive software and electric transportation, as well as simulation systems for robotics and traffic.