Visual-in-Visual: A Unified and Efficient Baseline for Image Restoration

Yuning Cui, *Graduate Student Member, IEEE*, Wenqi Ren, *Senior Member, IEEE*, Boxin Shi, *Senior Member, IEEE*, and Alois Knoll, *Fellow, IEEE*

***Abstract*—Recent years have witnessed remarkable progress in image restoration, yet achieving both high performance and efficiency remains a persistent challenge. To address this issue, we present VIVNet, a strong and efficient unified baseline designed to balance accuracy and practicality. Drawing inspiration from the high efficiency of the human visual system, VIVNet embeds a biologically inspired micro visual module into each block of a macro U-shaped vision architecture. This module mimics key perceptual processes such as retinal encoding, lateral inhibition, and high-order processing by combining lightweight depth-wise convolutions for multi-receptive-field feature extraction, a similarity-aware weighting mechanism to emphasize informative signals, and high-order interactions implemented via iterative element-wise multiplication to capture complex dependencies. This design enhances the model's representational capacity while maintaining computational efficiency. Unlike most existing methods that are limited to narrow task settings, we evaluate VIVNet across a wide range of scenarios, including general, all-in-one, and composite degradation tasks, as well as ultra-high-definition (UHD), underwater, medical, and remote sensing datasets. Extensive experiments show that VIVNet delivers competitive performance with high efficiency.**

***Index Terms*—Image restoration, all-in-one image restoration, composite degradation, unified network, efficient baseline, remote sensing, medical image processing, underwater image enhancement, ultra-high-definition image restoration, human vision**

I. Introduction

IMAGES captured under adverse weather conditions or in dynamic environments often suffer from degraded visibility, which can significantly impair the performance of downstream vision tasks such as object detection, depth estimation, and action recognition. In this context, image restoration, a long-standing and inherently ill-posed task, was originally employed to enhance the visibility of space imagery by mitigating various degradations. The rise of deep learning has driven substantial progress in the field, with learning-based methods significantly outperforming traditional approaches that depend on hand-crafted priors and strong assumptions [1].

Early deep learning-based methods primarily focused on *single-degradation* image restoration and have achieved remarkable performance on various tasks, including dehazing [2]–[4], deraining [5]–[7], and deblurring [8]–[10]. These methods perform well on a specific task; however, they struggle to be generalized to other tasks. Subsequently, some *general* frameworks have been developed to address multiple degradation types using the same or scaled architectures. Representative approaches in this category include CNN-based [11]–[13], Transformer-based [14]–[17], and Mamba-based [18]–[20] models. Nonetheless, users have to train separate model instances for different tasks, which limits their usage on resource-constrained platforms. To mitigate these issues, *all-in-one* methods have recently attracted significant attention in image restoration by addressing diverse degradations within a single unified model [21]–[26]. These methods employ various strategies to extract degradation-related signals from corrupted images, which are then used to guide the degradation-aware restoration process. Common strategies include contrastive learning, prompt learning, frequency priors, and contextual cues. However, each image considered in such methods is typically subject to only a single type of degradation. To handle images simultaneously impaired by multiple degradation types, recent studies have proposed methods for *composite degradation* image restoration [27]–[30].

While the above methods have considerably advanced image restoration, substantial room remains for improving performance across diverse settings without increasing computational cost. Meanwhile, in real-world applications, different scenarios impose varying demands on restoration algorithms. However, most approaches lack the generality to perform effectively across such diverse settings. In this work, we aim to develop a method with broader applicability, achieving strong performance over a wider spectrum of image restoration tasks.

The encoder–decoder architecture, particularly the U-shaped design [31], exhibits functional parallels to several key mechanisms of biological vision and can thus be regarded as a biologically inspired structure. During the encoding phase, for example, the network extracts multi-scale features through successive downsampling, mirroring the hierarchical processing in the human visual system, where visual information is progressively transformed from the retina to the primary visual cortex (V1) across multiple spatial scales [32]. Owing to its efficiency in hierarchical representation learning, this design has been widely adopted in image restoration. To achieve our two primary objectives, efficiency and universality, we draw inspiration from the human visual system, renowned for its remarkable adaptability and processing efficiency. As illustrated in Figure 1(a), we abstract it into three func-

Corresponding author: Wenqi Ren (renwq3@mail.sysu.edu.cn)

Yuning Cui 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, China, and also with the State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, 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.

The code is available at https://github.com/c-yn/VIVNet

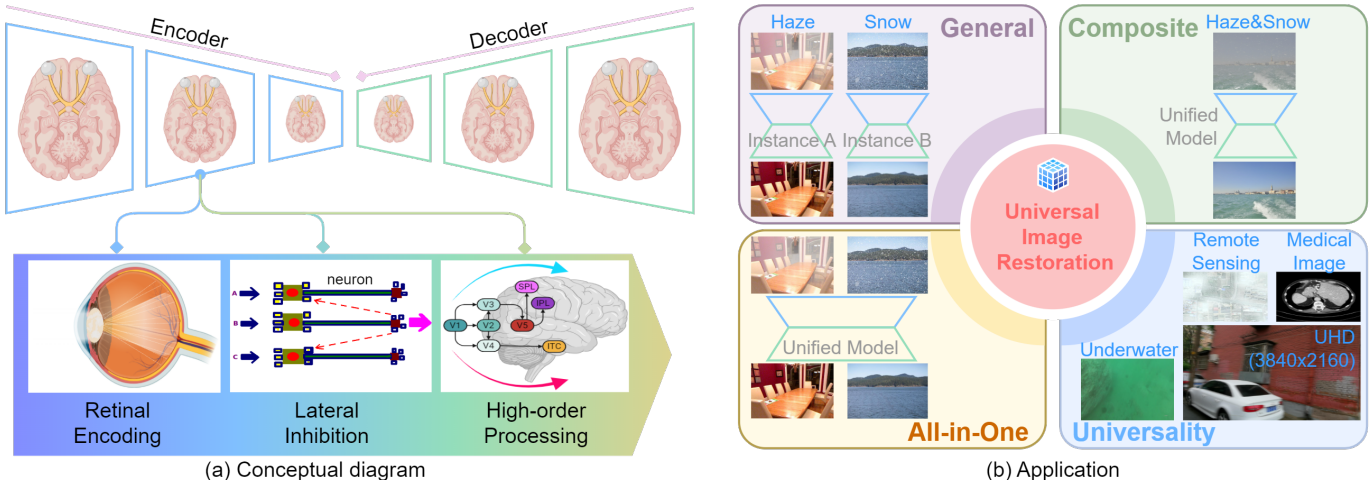


Fig. 1. (a) High-level conceptual diagram of the proposed visual-in-visual network. (b) Overview of the image restoration tasks supported by the method, covering various degradation settings and application domains.

tional stages: retinal encoding, lateral inhibition, and high-order processing. Specifically, retinal encoding captures visual cues at different spatial scales through the complementary roles of foveal and peripheral vision [33]; lateral inhibition enhances salient information by suppressing redundant or less informative neighboring responses [34]; and high-order processing captures complex dependencies through nonlinear interactions among multi-scale contextual features, enabling robust semantic understanding and abstraction [35].

Inspired by the three stages of human visual processing, we design a brain-inspired module comprising three components: encoding, similarity-aware weighting, and high-order interaction. To maintain efficiency, each component is implemented with lightweight yet effective operations. Concretely, the encoding stage employs depth-wise convolutions with varying kernel sizes to emulate the complementary roles of foveal and peripheral vision, thereby enabling feature extraction across multiple receptive fields. The similarity-aware weighting mechanism then enhances informative signals by generating adaptive weights based on the cosine similarity between extracted features. Finally, high-order interaction is realized through iterative element-wise multiplication, capturing nonlinear dependencies among features [36], [37]. At the architectural level, we embed the proposed module into a macro visual U-shaped framework, following the Network-in-Network [38] design philosophy, resulting in our Visual-in-Visual Network (VIVNet).

The evaluation of existing image restoration methods is commonly confined to a narrow range of datasets, potentially overlooking the complexity and variability of real-world conditions. In contrast, as shown in Figure 1(b), we perform a comprehensive evaluation of our model across diverse settings and tasks. This includes 12 datasets covering six single-degradation tasks for general-purpose image restoration, two configurations for all-in-one restoration, and two composite degradation tasks. We further evaluate its performance on four UHD image restoration tasks. Beyond natural images, we extend the evaluation to domain-specific applications using seven datasets spanning medical imaging, remote sensing, and underwater image enhancement. As illustrated in Figure 2, the proposed network consistently outperforms recent state-of-the-art methods while maintaining high parameter efficiency. The main contributions of this work are summarized as follows:

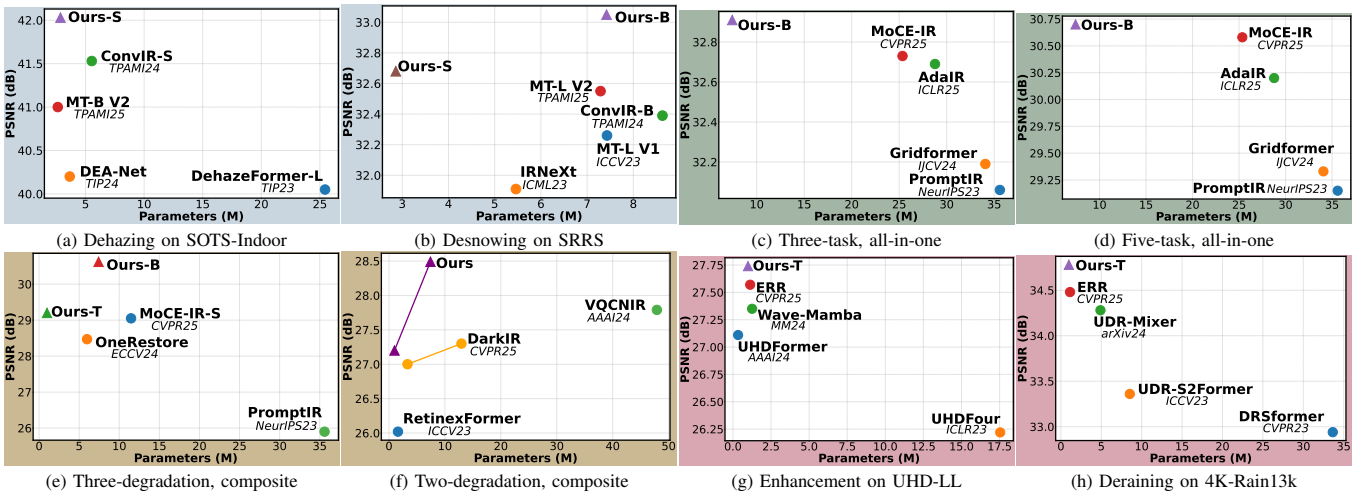
- We introduce VIVNet, a Visual-in-Visual Network that integrates a lightweight, brain-inspired module designed to emulate key stages of human visual perception.
- We propose a similarity-aware weighting mechanism that adaptively emphasizes informative features by deriving channel-wise weights from the cosine similarity of features. In addition, high-order feature interactions are modeled through iterative element-wise operations.
- Extensive experiments across general-purpose, all-in-one, composite degradation, UHD, and domain-specific restoration tasks demonstrate that VIVNet achieves state-of-the-art performance while maintaining high computational efficiency.

II. Related Work

A. Image restoration framework

As a fundamental task in computer vision, image restoration aims to recover a sharp image from the degraded input. Traditional methods primarily rely on carefully designed priors to constrain the solution space [39]. While such priors are effective under certain conditions, their statistical nature limits their ability to generalize to complex or atypical scenes.

The rapid development of deep learning has led to a proliferation of convolutional networks that substantially outperform traditional approaches across diverse image restoration tasks, including dehazing [2]–[4], deraining [5]–[7], and deblurring [9], [10], [40]. Incorporating advanced functional modules, such as multi-stage paradigms, dilated convolutions, dynamic convolutions, and various attention mechanisms, has further enhanced restoration performance. For instance, FFA-Net [2] employs channel and pixel attention for image dehazing, while MPRNet [11] leverages spatial attention to control



(a) Dehazing on SOTS-Indoor (b) Desnowing on SRRS (c) Three-task, all-in-one (d) Five-task, all-in-one

(e) Three-degradation, composite (f) Two-degradation, composite (g) Enhancement on UHD-LL (h) Deraining on 4K-Rain13k

Fig. 2. Comparison between state-of-the-art algorithms and our methods in terms of parameter count and PSNR across various image restoration settings, including general-purpose, all-in-one, composite degradation, and ultra-high-definition.

information flow between stages for general-purpose restoration. Nevertheless, the inherent limitation of the convolution operator, namely its local connectivity, restricts the ability of CNN-based methods to model large-scale dependencies, which are crucial for addressing severe degradations.

This limitation has been largely addressed by the introduction of Transformer architectures [16], [17], [41], [42]. However, the canonical self-attention mechanism has quadratic computational complexity, making it impractical for image restoration, which often involves high-resolution images. To improve the efficiency of Transformer-based frameworks, numerous strategies have been explored. For example, some methods restrict the self-attention operation to local regions [15], [16], [43], [44], while others apply self-attention across channels rather than in the spatial dimension [14], [45].

Lately, Mamba-based frameworks have further improved the efficiency of image restoration methods by achieving linear complexity. They are primarily built upon advanced scanning strategies or the incorporation of complementary local information [18]–[20], [46]–[48]. For example, MambaLLIE [49] integrates local invariance into the state space model, while EAMamba [50] proposes an all-around scanning strategy that combines two-dimensional and diagonal scanning.

In contrast to the aforementioned methods that employ advanced architectures for image restoration, we propose a visual-in-visual framework that emulates the mechanisms of the human visual system using lightweight operators.

B. Multi-task image restoration

In this study, we classify both all-in-one and composite degradation image restoration approaches as multi-task methods. Different from the single-degradation and general-purpose image restoration algorithms, multi-task approaches can handle multiple degradations with a single model, distributed across different images or existing simultaneously in a single image [51]. The widely used strategy for multi-task learning is to learn the degradation-related information from input images, which is then used to guide the image restoration process. For instance, AirNet [21] learns degradation representations from corrupted images through contrastive learning, whose performance heavily depends on the accurate selection of positive and negative pairs. InstructIR [26] is the first to utilize human-written instructions to guide the image restoration process. Wu *et al.* [52] introduce a reweighting strategy to harmonize optimization across multiple degradation tasks and mitigate conflicts in multitask learning. Zamfir *et al.* [29] design complexity experts, specialized units with varying receptive fields and computational demands, to improve task-discriminative learning and accelerate inference by bypassing irrelevant experts. Low-rank designs have also been explored in this domain [53], [54]. More recently, Guo *et al.* [28] develop a scene descriptor-guided cross-attention block within a Transformer-based model, enabling adaptable restoration for composite degradations.

Although not specifically designed for multi-task image restoration, our method achieves strong performance on both all-in-one and composite degradation datasets, demonstrating its powerful representation learning capability.

C. Ultra-high-definition image restoration

With the rapid advancement of imaging technology, UHD images with high pixel density and resolution (*e.g.*, 3840×2160 or higher) have become increasingly prevalent in daily life [55]. Existing UHD image restoration methods can be broadly categorized into three architectural types: single-branch, dual-branch, and U-shaped. In single-branch approaches, the UHD image is first downsampled to a lower-resolution space to improve efficiency during feature extraction, and the refined features are subsequently upsampled to the original resolution [56]. Dual-branch methods augment the downsampled branch with a high-resolution branch and perform feature interactions between the two branches [57]–[59]. The third category employs a U-shaped architecture to capture hierarchical representations [60], [61]. Our model adopts the U-shaped design to ensure compatibility with diverse image restoration tasks and integrates a visual-in-visual mechanism to enhance both effectiveness and efficiency.

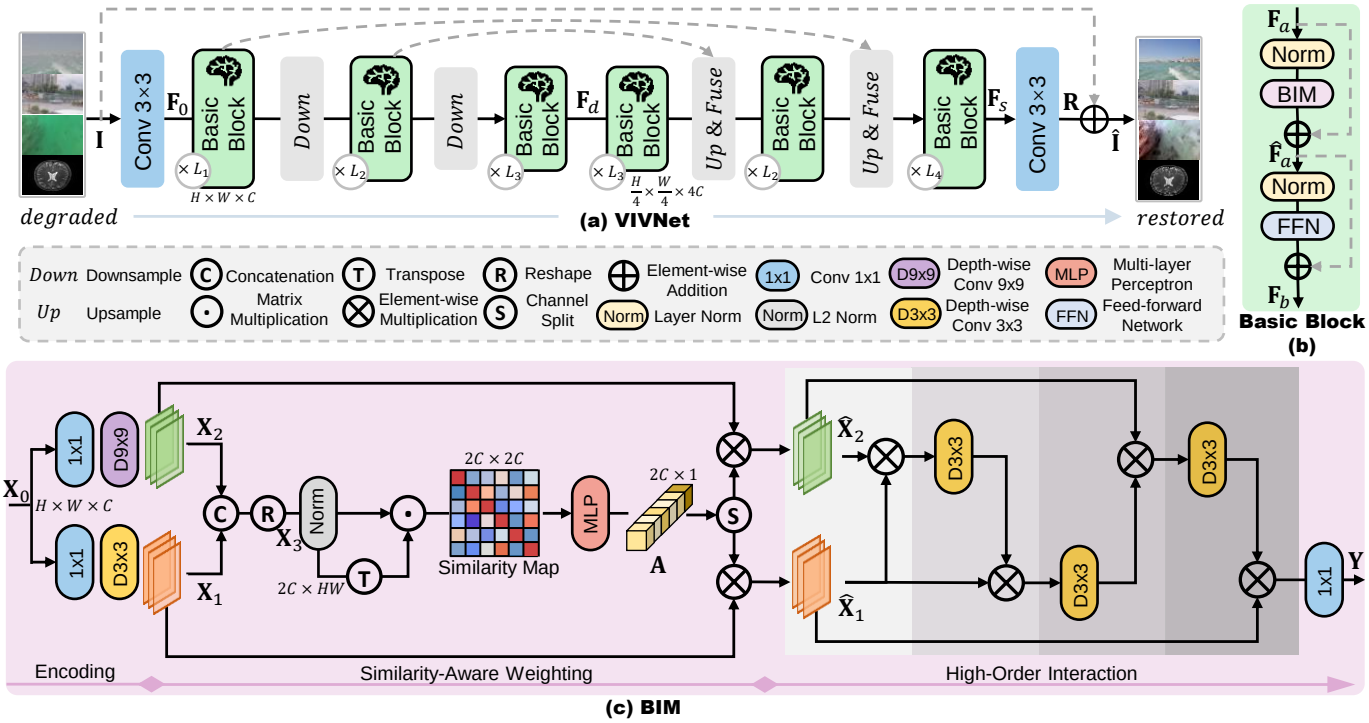


Fig. 3. **(a)** Architecture of the Visual-in-Visual Network (VIVNet) for universal image restoration, featuring a U-shaped macro visual design and micro visual basic blocks at each stage. **(b)** Each basic block primarily comprises the proposed Brain-Inspired Module (BIM) and a feed-forward network (FFN) [14]. **(c)** The BIM includes three processing stages: (1) the encoding stage captures visual cues at multiple spatial scales by emulating foveal and peripheral vision mechanisms; (2) the similarity-aware weighting mechanism assigns weights based on a cosine similarity map to emphasize informative features and suppress redundancy; and (3) the iterative use of element-wise multiplication combined with depth-wise convolutions to model high-order interactions. Overall, the BIM leverages lightweight operations to mimic human visual processing while maintaining high efficiency.

III. Methodology

A schematic of the VIVNet is shown in Figure 3. In this section, we first present the overall pipeline of VIVNet for universal image restoration. We then describe the Brain-Inspired Module (BIM) in detail, which forms the core of the architecture and comprises three key stages: (1) the encoding stage captures visual cues at multiple spatial scales; (2) the similarity-aware mechanism enhances informative signals by generating weights from the cosine similarity map; and (3) high-order processing models non-linear interactions among multi-scale features through a combination of element-wise multiplication and depth-wise convolutions.

A. Overall pipeline

The overall pipeline of VIVNet is shown in Figure 3(a). At the macro level, VIVNet adopts a plain U-shaped architecture to efficiently extract hierarchical representations, without incorporating additional techniques such as multi-input or multi-output strategies [8], [62].

Given a degraded image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the proposed network first applies a 3×3 convolutional layer to extract low-level features $\mathbf{F}_0 \in \mathbb{R}^{H\times W\times C}$. These features are then processed by a three-scale encoder, where the spatial resolution is progressively reduced and the channel dimension expanded, producing deep features $\mathbf{F}_d \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 4C}$. Each scale comprises multiple basic blocks, whose architectural details are shown in Figure 3(b). Given input features $\mathbf{F}_a$, the computation within a basic block is formally defined as:

$$\hat{\mathbf{F}}_a = \mathcal{B}(\mathcal{N}(\mathbf{F}_a)) + \mathbf{F}_a, \tag{1}$$

$$\mathbf{F}_b = \mathcal{F}(\mathcal{N}(\hat{\mathbf{F}}_a)) + \hat{\mathbf{F}}_a, \tag{2}$$

where $\hat{\mathbf{F}}_a$ and $\mathbf{F}_b$ denote the intermediate and output features, respectively, while $\mathcal{N}$, $\mathcal{B}$, and $\mathcal{F}$ represent layer normalization, the proposed BIM, and the feed-forward network (FFN), respectively. Since our primary focus is on designing the brain-inspired mechanism, we directly adopt the FFN module from Restormer [14] to streamline the overall architecture.

Subsequently, $\mathbf{F}_d$ is passed through a three-scale decoder to obtain sharp features. In this stage, the spatial resolution is progressively restored to the original size, producing $\mathbf{F}_s \in \mathbb{R}^{H\times W\times C}$. Residual connections are used to concatenate the corresponding encoder and decoder features, followed by a 1×1 convolution to adjust the channel dimension. A 3×3 convolution is then applied to generate the residual clean image $\mathbf{R} \in \mathbb{R}^{H\times W\times 3}$. Finally, the restored image is computed as $\hat{\mathbf{I}} = \mathbf{I} + \mathbf{R}$. The network parameters are optimized using the L_1 loss [8], [62]:

$$\mathcal{L}_s(\hat{\mathbf{I}}, \mathbf{I}^*) = \|\hat{\mathbf{I}} - \mathbf{I}^*\|_1, \tag{3}$$

$$\mathcal{L}_f(\hat{\mathbf{I}}, \mathbf{I}^*) = \|\mathrm{FFT}(\hat{\mathbf{I}}) - \mathrm{FFT}(\mathbf{I}^*)\|_1, \tag{4}$$

where $\mathcal{L}_s(\cdot)$ and $\mathcal{L}_f(\cdot)$ denote the losses in the spatial and frequency domains, respectively, and $\mathbf{I}^*$ represents the ground-truth image. FFT refers to the Fast Fourier Transform. The

final loss is defined as a weighted combination of the two components: $\mathcal{L} = \mathcal{L}_s(\hat{\mathbf{I}}, \mathbf{I}^*) + 0.1\mathcal{L}_f(\hat{\mathbf{I}}, \mathbf{I}^*)$.

B. Brain-Inspired Module (BIM)

Efficiency and effectiveness are critical for the practical deployment of image restoration algorithms. Although recent approaches have achieved notable progress by employing advanced architectures such as Transformer and Mamba, there remains substantial scope for improving performance without sacrificing efficiency. Furthermore, many existing methods are limited to a narrow set of tasks, overlooking the complexity and diversity of real-world scenarios.

To address these limitations, we take inspiration from nature, specifically, the efficiency and universality of the human visual system, rather than resorting to increasingly complex architectures or sophisticated attention mechanisms. We examine three key processes in biological vision: retinal encoding, lateral inhibition, and high-order feature integration. Guided by these principles, we design the BIM, comprising three components that conceptually correspond to these functions, as illustrated in Figure 3(c). Leveraging lightweight and structurally simple operators, the BIM achieves a strong representational capacity while maintaining high computational efficiency. We next detail the individual components of BIM.

1) Multi-scale encoding: The initial stage of the human visual system involves perceiving the environment through the retina, where foveal and peripheral vision operate at distinct spatial scales. This multi-scale processing allows receptive fields to adapt dynamically to varying stimuli, thereby supporting robust and high-fidelity semantic perception. To replicate this mechanism in our model, we employ lightweight depth-wise convolutions with different kernel sizes to capture receptive fields of varying extents while keeping computational overhead low. Given input features $\mathbf{X}_0 \in \mathbb{R}^{H\times W\times C}$, the multi-scale encoding stage is formulated as:

$$\mathbf{X}_1 = \mathrm{Conv}^d_{3\times3}(\mathrm{Conv}_{1\times1}(\mathbf{X}_0)), \tag{5}$$

$$\mathbf{X}_2 = \mathrm{Conv}^d_{9\times9}(\mathrm{Conv}_{1\times1}(\mathbf{X}_0)), \tag{6}$$

where $\mathrm{Conv}_{1\times1}$ denotes a standard convolution with a 1×1 kernel, and $\mathrm{Conv}^d_{3\times3}$ denotes a depth-wise convolution with a 3×3 kernel. The resulting features $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{H\times W\times C}$ encode visual cues at different receptive fields.

2) Similarity-aware weighting mechanism: In the human visual system, visual cues are not immediately subjected to high-level processing; rather, they undergo pre-processing through mechanisms such as lateral inhibition, which enhance informative signals while suppressing redundancy. Inspired by this mechanism, we employ cosine similarity to measure the relative similarity between extracted features. The resulting similarity map is then used to compute attention weights that emphasize salient responses and attenuate redundant or repetitive patterns. We next describe the similarity computation and subsequent weight generation in detail.

Similarity measurement. We first revisit the formulation of cosine similarity before describing its application in our framework. Given two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, the cosine similarity is defined as:

$$\mathrm{CosSim}(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b}\rangle}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2}, \tag{7}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. To compute the inter-channel similarity between the features $\mathbf{X}_1$ and $\mathbf{X}_2$, we first concatenate them and reshape the result, obtaining $\mathbf{X}_3 \in \mathbb{R}^{2C\times HW}$. Following Eq. (7), the cosine similarity is then computed as:

$$\mathbf{S}_{ij} = \frac{\langle (\mathbf{X}_3)_i, (\mathbf{X}_3)_j\rangle}{\|(\mathbf{X}_3)_i\|_2\|(\mathbf{X}_3)_j\|_2}, \quad \forall i,j \in \{1,\dots,2C\}, \tag{8}$$

where $\mathbf{S}$ denotes the channel-wise similarity matrix computed via cosine similarity, with each row representing the affinities between a given channel and all channels. This formulation allows the network to capture structural relationships and dependencies across channels.

Weight generation. Based on the similarity matrix $\mathbf{S}$, we apply a lightweight multi-layer perceptron (MLP) to generate attention weights, enabling the model to selectively emphasize informative channels while suppressing less relevant ones. This process is formally defined as:

$$\mathbf{A} = W_2\sigma(W_1\mathbf{S}), \tag{9}$$

where $\mathbf{A} \in \mathbb{R}^{2C\times1}$ denotes the computed channel-wise weights. W_1 and W_2 are linear layers that progressively reduce the channel dimension from $2C$ to $0.6\times2C$, and finally to 1. σ represents the LeakyReLU activation function.

Finally, $\mathbf{A}$ is split and applied to the original inputs to produce the pre-processed features $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$, which are then used for subsequent high-order interactions.

3) High-order interaction: After redundant information is suppressed during early visual processing, the selected visual cues are transmitted to the cerebral cortex, where complex neuronal interactions support robust semantic understanding. Inspired by this process, we simulate high-level interactions by iteratively applying element-wise multiplication [36] and depth-wise convolutions to $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$.

More specifically, we first compute the first-order interaction by performing element-wise multiplication between the two input features. The second-order interaction is achieved by applying a 3×3 depth-wise convolution to the resulting features, followed by element-wise multiplication with one of the original inputs. This procedure is iteratively repeated to capture high-order interactions between the two features. Finally, the output is generated through a 1×1 convolutional layer. The computation in the high-order interaction stage can be formally expressed as:

$$\mathbf{Z}^{(1)} = \hat{\mathbf{X}}_1 \otimes \hat{\mathbf{X}}_2, \tag{10}$$

$$\mathbf{Z}^{(t)} = \mathrm{Conv}^d_{3\times3}\big(\mathbf{Z}^{(t-1)}\big) \otimes \hat{\mathbf{X}}_{(t \bmod 2)+1}, \quad t = 2,\dots,N \tag{11}$$

$$\mathbf{Y} = \mathrm{Conv}_{1\times1}\big(\mathbf{Z}^{(N)}\big), \tag{12}$$

where $\mathbf{Z}$ and $\mathbf{Y}$ denote the intermediate and final results, respectively, and $\otimes$ represents element-wise multiplication. The operator mod denotes the modulo operation, such that

TABLE I
ARCHITECTURAL CONFIGURATIONS OF FOUR VARIANTS OF THE PROPOSED NETWORK.

Variant	Num. of Blocks	Num. of Channels	Params	FLOPs
Ours-T (*Tiny*)	[1,1,1,1,1,5]	[32,64,128,128,64,32,32]	0.99M	14.18G
Ours-S (*Small*)	[2,3,4,4,3,6]	[32,64,128,128,64,32,32]	2.86M	25.97G
Ours-B (*Base*)	[3,3,5,5,3,7]	[48,96,192,192,96,48,48]	7.42M	63.40G
Ours-L (*Large*)	[6,6,12,12,6,10]	[48,96,192,192,96,48,48]	16.48M	117.74G

$\hat{\mathbf{X}}_{(t \bmod 2)+1}$ alternates between $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_2$ as the interaction order t increases. As a result, the high-order spatial interaction enhances robust scene understanding (see Figure 13).

IV. EXPERIMENTAL RESULTS

To demonstrate the generality and effectiveness of the proposed VIVNet, we conduct extensive experiments across a wide range of image restoration tasks, spanning both synthetic and real-world datasets. This section presents results for: **(a)** regular tracks, including general-purpose, all-in-one, and composite degradation image restoration; **(b)** UHD image restoration; and **(c)** domain-specific tasks, such as medical, underwater, and remote sensing image restoration. Finally, we present the results of the ablation study and failure cases.

Due to the varying complexity of the benchmark datasets, and following previous work [62], [66], we adopt different versions of our model, namely *tiny* (T), *small* (S), *base* (B), and *large* (L), for different datasets. This ensures that our model maintains computational overhead comparable to that of competing methods on each individual dataset, thereby enabling a fair comparison. A summary of the model specifications is provided in Table I. FLOPs are measured on $3 \times 256 \times 256$ patches. Further details on the datasets, implementation, and additional visual results are included in the supplementary material. Importantly, our training and evaluation protocols strictly follow prior works, without introducing any extra techniques, to guarantee fairness in comparison. The best and second-best quality results in the tables are highlighted in magenta and blue, respectively.

A. General image restoration

Following prior general-purpose image restoration methods [11], [14], [62], [66], we train separate model instances for each dataset in degradation-specific tasks, including desnowing, deraining, low-light image enhancement, defocus deblurring, motion deblurring, and dehazing.

1) Image desnowing: We evaluate our model on two image desnowing datasets: SRRS [63] and Snow100K [64]. Table II shows that our base version consistently surpasses the Transformer-based MT-L V2 [66] across all quality metrics, while reducing computational complexity by 26% and maintaining a comparable parameter count, highlighting its effectiveness and efficiency. Compared to ConvIR-B [62], which integrates multiple complex convolutional modules to boost performance, our model delivers notable performance gains with a simpler design, resulting in lower FLOPs and fewer parameters. We further report the results of our small version on these two datasets. As expected, this variant attains lower quantitative scores than the base model due to the reduced computational capacity. Nevertheless, it still surpasses competing approaches on both datasets while requiring substantially lower computational overhead.

2) Image deraining: We evaluate our model on the widely used synthetic DID-Data [67] and real-world SPA-Data [68] datasets for image deraining. Following prior works [69], [70], [72], evaluation metrics are computed on the Y channel of the YCbCr color space. As shown in Table III, our approach achieves the best performance on both datasets compared with recent state-of-the-art algorithms. Notably, it outperforms the specialized deraining method FourierMamba [72] by 0.36 dB in PSNR on the real-world dataset while requiring only 42% of its parameters. We also extend the evaluation to raindrop removal using the AGAN-Data [73] dataset, where our method achieves substantial improvements over previous state-of-the-art approaches, as reported in Table IV.

3) Low-light image enhancement: For this task, we evaluate our model on the widely adopted LOL-v2-s [78] dataset. Table V shows that our method achieves 26.01 dB PSNR, exceeding the recent Mamba-based MambaLLIE [49] and MambaIR [18] by 0.14 dB and 0.46 dB, respectively. It also outperforms the strong Transformer-based Retinexformer [80] by 0.34 dB in PSNR. These results demonstrate the strong effectiveness of our approach for enhancing low-light images.

4) Image deblurring: We evaluate VIVNet on both defocus deblurring and motion deblurring tasks. For single-image defocus deblurring, we evaluate it on the DPDD [81] dataset. As shown in Table VI, our method achieves the best performance on most metrics. In the combined categories, it surpasses the recent state-of-the-art algorithm [83] by 0.16 dB in PSNR and 0.006 in SSIM. Notably, compared with the more recent RDDM [84], a task-specific approach, our model delivers a substantial improvement of 0.31 dB in PSNR, underscoring its effectiveness for defocus deblurring.

Table VII presents the results on GoPro [85] for motion deblurring. Our method surpasses the recent Mamba-based EAMamba [50] by 0.1 dB in PSNR while requiring 35% fewer parameters. Compared to MT-XL V2 [66], it achieves a 0.44 dB PSNR improvement with a comparable parameter count.

5) Image dehazing: We conduct image dehazing experiments on two synthetic datasets (SOTS-Indoor [94] and Haze4k [91]) and two real-world datasets (Dense-Haze [88] and NH-HAZE [89]). As shown in Table VIII, our model outperforms the second-best method, ConvIR-S [62], on the SOTS-Indoor dataset by 0.5 dB in PSNR, while requiring only half the parameters and 38% less computational complexity. Furthermore, it exhibits strong robustness on real-world datasets, achieving the best performance on most metrics.

Following [62], [66], [93], we also evaluate on the more realistically synthesized Haze4k dataset. Table IX shows that our base model consistently outperforms recent strong competitors while incurring lower computational overhead. In addition, we compare the results obtained by different versions of our model. Our findings indicate a clear trend: models with higher computational overhead achieve better performance.

TABLE II
QUANTITATIVE COMPARISONS ON THE SRRS [63] AND SNOW100K [64] DATASETS FOR IMAGE DESNOWING.

Datasets	Metrics	FocalNet [12]	IRNeXt [13]	ConvIR-B [62]	MT-L V1 [65]	MT-L V2 [66]	**Ours-S**	**Ours-B**
SRRS	PSNR	31.34	31.91	32.39	32.26	32.55	32.68	33.05
	SSIM	0.98	0.98	0.98	0.98	0.98	0.98	0.98
Snow100K	PSNR	33.53	33.61	33.92	33.79	34.01	34.25	34.52
	SSIM	0.95	0.95	0.96	0.95	0.96	0.96	0.96
Overhead	Params	3.74M	5.46M	8.63M	7.43M	7.29M	2.86M	7.42M
	FLOPs	30.6G	42.1G	71.2G	88.1G	86.0G	25.97G	63.4G

TABLE III
DERAINING RESULTS ON THE SYNTHETIC DID-DATA [67] AND REAL-WORLD SPA-DATA [68] DATASETS.

Method	DID-Data PSNR	DID-Data SSIM	SPA-Data PSNR	SPA-Data SSIM	Params
Restormer [14]	35.29	0.9641	47.98	0.9921	26.13M
DRSformer [69]	35.35	0.9646	48.54	0.9924	33.65M
NeRD-Rain-S [70]	35.36	0.9647	48.90	0.9936	10.53M
FADformer [71]	35.48	0.9657	49.21	0.9934	6.96M
FourierMamba [72]	35.49	0.9659	49.18	0.9931	17.62M
Ours-B	35.57	0.9663	49.54	0.9938	7.42M

TABLE IV
DERAINING RESULTS ON AGAN-DATA [73].

Method	IDT [74]	MAXIM [75]	AWRCP [76]	FPro [77]	**Ours-B**
PSNR	31.87	31.87	31.93	31.96	33.51
SSIM	0.931	0.935	0.931	0.937	0.947

TABLE V
ENHANCEMENT RESULTS ON LOL-V2-S [78].

Method	Retinexformer [79]	MambaIR [80]	MambaLLIE [18]	**Ours-T** [49]
PSNR	25.67	25.55	25.87	26.01
SSIM	0.930	0.929	0.940	0.948

TABLE VI
SINGLE-IMAGE DEFOCUS DEBLURRING COMPARISONS ON THE DPDD [81] TEST SET (CONTAINING 37 INDOOR AND 39 OUTDOOR SCENES).

Method	Indoor Scenes PSNR↑	SSIM↑	MAE↓	LPIPS↓	Outdoor Scenes PSNR	SSIM	MAE	LPIPS	Combined PSNR	SSIM	MAE	LPIPS
DPDNet [81]	26.54	0.816	0.031	0.239	22.25	0.682	0.056	0.313	24.34	0.747	0.044	0.277
KPAC [10]	27.97	0.852	0.026	0.182	22.62	0.701	0.053	0.269	25.22	0.774	0.040	0.227
IFAN [9]	28.11	0.861	0.026	0.179	22.76	0.720	0.052	0.254	25.37	0.789	0.039	0.217
DRBNet [40]	28.49	0.863	0.038	0.141	23.10	0.724	0.063	0.222	25.73	0.791	0.051	0.183
Restormer [14]	28.87	0.882	0.025	0.145	23.24	0.743	0.050	0.209	25.98	0.811	0.038	0.178
NRKNet [82]		-				-			26.11	0.810	-	0.210
SFHformer [83]	28.95	0.874	0.024	0.182	23.44	0.743	0.049	0.260	26.12	0.807	0.037	0.222
RDDM [84]		-				-			25.97	0.811	0.037	0.166
Ours-L	29.22	0.883	0.024	0.156	23.49	0.746	0.049	0.221	26.28	0.813	0.037	0.189

TABLE VII
QUANTITATIVE RESULTS ON THE GOPRO [85] DATASET FOR MOTION DEBLURRING.

Method	PSNR	SSIM	Params
Restormer [14]	32.92	0.961	26.1M
Stripformer [43]	33.08	0.962	20M
FSNet [86]	33.29	0.963	13.28M
IRNeXt [13]	33.16	0.962	13.21M
ConvIR-L [62]	33.28	0.963	14.83M
MT-XL V2 [66]	33.24	0.963	16.26M
EAMamba [50]	33.58	0.966	25.3M
Ours-L	33.68	0.967	16.48M

TABLE VIII
QUANTITATIVE COMPARISONS ON THE SYNTHETIC SOTS-INDOOR DATASET [87] AND TWO REAL-WORLD DATASETS, DENSE-HAZE [88] AND NH-HAZE [89], FOR THE IMAGE DEHAZING TASK.

Method	SOTS-Indoor PSNR	SOTS-Indoor SSIM	Dense-Haze PSNR	Dense-Haze SSIM	NH-HAZE PSNR	NH-HAZE SSIM	Overhead Params	Overhead FLOPs
Dehamer [41]	36.63	0.988	16.62	0.560	20.66	0.684	132.50M	60.3G
DehazeFormer-L [44]	40.05	0.996	-	-	-	-	25.44M	279.7
MT-B V1 [65]	40.71	0.992	16.66	0.560	20.43	0.688	2.68M	38.5G
DEA-Net [90]	40.20	0.993	-	-	-	-	3.65M	32.23G
ConvIR-S [62]	41.53	0.994	17.45	0.608	20.65	0.692	5.53M	42.1G
MT-B V2 [66]	41.00	0.993	16.95	0.621	20.73	0.703	2.63M	37.7G
Ours-S	42.03	0.998	17.26	0.632	21.01	0.706	2.86M	25.97G

TABLE IX
QUANTITATIVE COMPARISONS ON THE HAZE4K [91] DATASET FOR IMAGE DEHAZING.

Method	FFA-Net [2]	PMNet [92]	FSNet [93]	ConvIR-S [62]	ConvIR-B [62]	ConvIR-L [62]	MT-L V1 [65]	MT-L V2 [66]	**Ours-T**	**Ours-S**	**Ours-B**	**Ours-L**
PSNR	26.96	33.49	34.12	33.36	34.15	34.50	34.47	34.92	31.60	33.80	35.27	36.43
SSIM	0.95	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99	0.99	0.99
Params	4.46M	18.90M	13.28M	5.53M	8.63M	14.83M	7.43M	7.29M	0.99M	2.86M	7.42M	16.48M
FLOPs	287.8G	81.1G	110.5G	42.1G	71.2G	129.9G	88.1G	86G	14.18G	25.97G	63.4G	117.74G

6) Visual results: Figure 4 shows the visual results for dehazing. Our method exhibits superior brightness perception, producing results closer to the reference images, particularly in the sky regions. The error heatmaps, comparing the restored images with the ground truths, shown in the bottom-right corners, further highlight the performance advantage.

The visual results of deraining are presented in Figure 5. Compared with the recent competing method [71], our model is more effective at removing rainy degradations of varying sizes. In the first example, FADformer even produces a lower PSNR score than the input rainy image, whereas our result remains much more faithful to the reference. Notably, in the third, extremely challenging case, FADformer fails to produce a viewable image, while our model is able to recover the overall structure and outline.

In addition, the deblurring results in Figure 6 show that our method generates a sharper reconstruction from the challenging blurry input compared with competing approaches.

B. All-in-one image restoration

Following [25], [29], we evaluate our model for all-in-one image restoration under two settings: three-task and five-task,

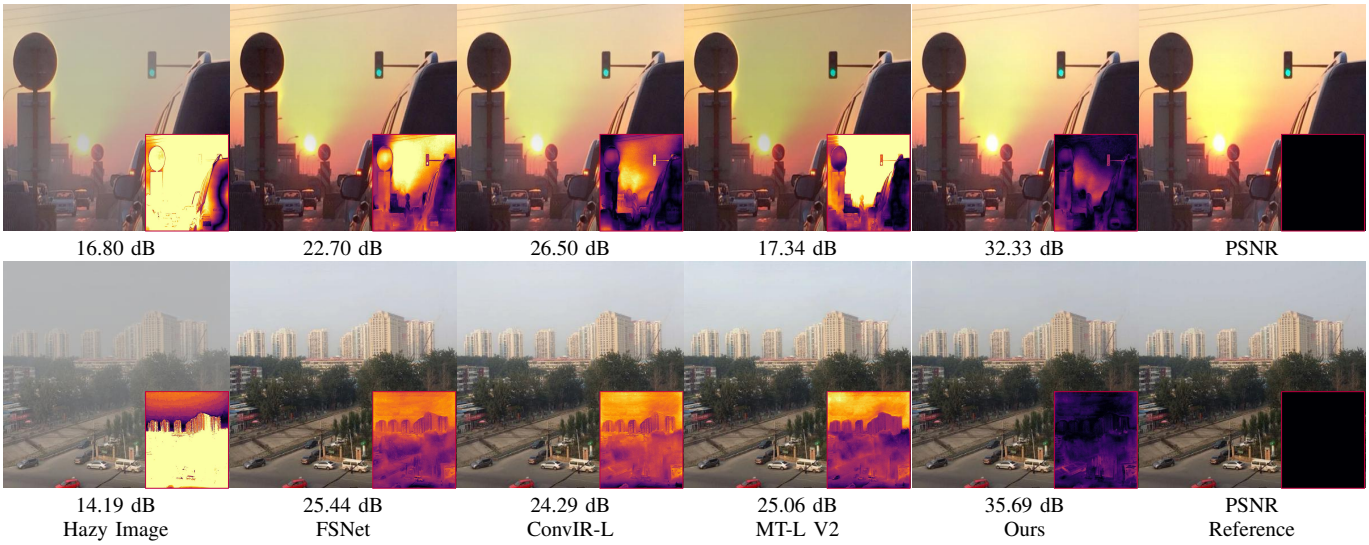


Fig. 4. Visual results on Haze4k [91] for dehazing. Our model exhibits superior brightness perception, producing results that more closely match the reference images. Specifically, in the first example, the sky regions appear brighter than those of the competitors, and in the second, darker, both consistent with the ground truth. Error heatmaps between restored and reference images are provided, with brighter pixels indicating larger errors.

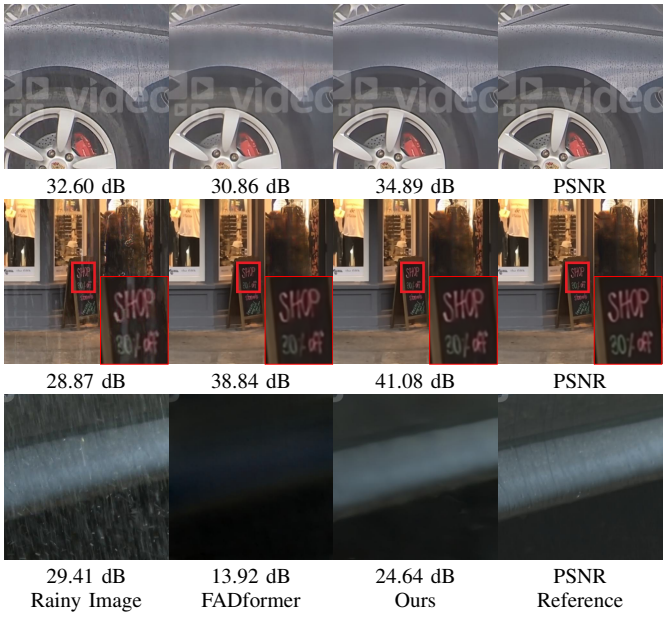


Fig. 5. Visual results on the real-world SPA-Data [68] dataset for deraining.

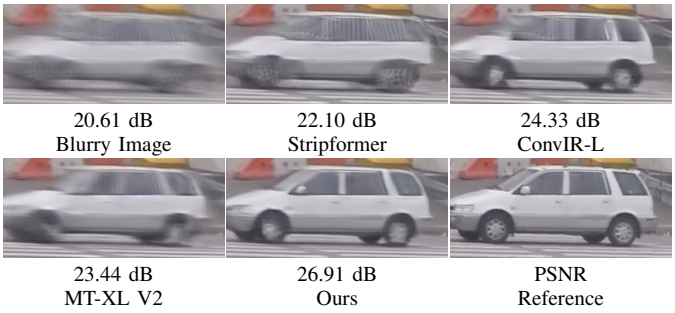


Fig. 6. Visual comparison of motion deblurring results on GoPro [85].

where the training datasets are collected from three and five degradation types, respectively. Each image contains only one type of degradation, and after training, a single unified model can handle multiple degradation types.

1) Three-task setting: For this setting, we conduct experiments on datasets derived from image denoising, deraining, and dehazing tasks. Quantitative comparisons are reported in Table X. The proposed model achieves the best results on both the dehazing and denoising tasks. Compared to the recent MoCE-IR [29], which adopts a mixture-of-experts strategy for improved efficiency, our approach achieves an average PSNR gain of 0.18 dB across all degradation types and a notable improvement of 0.85 dB on the dehazing task. This superior performance is achieved with only 29% of the parameters, highlighting the efficiency of our design.

2) Five-task setting: We further evaluate our model under a more challenging five-task setting by additionally including motion deblurring and low-light image enhancement. As shown in Table XI, our method achieves the best results on most datasets. When averaged across all datasets, it surpasses the small version of MoCE-IR [29] by 0.62 dB in PSNR. Even compared with its normal version, our model achieves a 0.12 dB improvement while reducing parameters by 71%. Furthermore, it outperforms the recent AdaIR [25], which leverages frequency priors to boost performance, on all metrics while using only 26% of the parameters.

The visual results on the real LOLv1 [97] dataset are shown in Figure 7. Our model is able to recover richer structural and textural details from the challenging low-light images compared with the recent frequency-based AdaIR [25], producing clearer and more visually faithful restorations.

3) Generalization evaluation: To assess the generalization capability of our model, we follow [25], [26] and directly apply the model pre-trained under the five-task setting to two additional denoising datasets: Urban100 [98] and Kodak24 [99]. Table XII shows that our model attains the highest scores on both datasets, outperforming strong competitors. In particular, under the challenging noise level of 50 on

TABLE X
QUANTITATIVE RESULTS FOR THE ALL-IN-ONE THREE-TASK SETTING. THE MODEL IS TRAINED ON A COMPOUND DATASET COMBINING THREE DEGRADATION TYPES. σ DENOTES THE NOISE LEVEL.

Method	Venue	Params	*Dehazing* SOTS		*Deraining* Rain100L		*Denoising* BSD68$_{\sigma=15}$		*Denoising* BSD68$_{\sigma=25}$		*Denoising* BSD68$_{\sigma=50}$		Average	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
AirNet [21]	CVPR22	8.93M	27.94	0.962	34.90	0.968	33.92	0.933	31.26	0.888	28.00	0.797	31.20	0.910
PromptIR [22]	NeurIPS23	35.6M	30.58	0.974	36.37	0.972	33.98	0.933	31.31	0.888	28.06	0.799	32.06	0.913
Art$_{\text{PromptIR}}$ [52]	MM24	33M	30.83	0.979	37.94	0.982	34.06	0.934	31.42	0.891	28.14	0.801	32.49	0.917
Gridformer [95]	IJCV24	34.07M	30.37	0.970	37.15	0.972	33.93	0.931	31.37	0.887	28.11	0.801	32.19	0.912
AdaIR [25]	ICLR25	28.77M	31.06	0.980	38.64	0.983	34.12	0.935	31.45	0.892	28.19	0.802	32.69	0.918
MoCE-IR [29]	CVPR25	25.35M	31.34	0.979	38.57	0.984	34.11	0.932	31.45	0.888	28.18	0.800	32.73	0.917
Ours-B	Ours	7.42M	32.19	0.982	38.47	0.983	34.16	0.936	31.50	0.893	28.24	0.806	32.91	0.920

TABLE XI
QUANTITATIVE RESULTS FOR THE ALL-IN-ONE FIVE-TASK SETTING.

Method	Venue	Params	*Dehazing* SOTS		*Deraining* Rain100L		*Denoising* BSD68$_{\sigma=25}$		*Deblurring* GoPro		*Low-Light* LOLv1		Average	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
IDR [96]	CVPR23	15.34M	25.24	0.943	35.63	0.965	31.60	0.887	27.87	0.846	21.34	0.826	28.34	0.893
PromptIR [22]	NeurIPS23	35.6M	26.54	0.949	36.37	0.970	31.47	0.886	28.71	0.881	22.68	0.832	29.15	0.904
Gridformer [95]	IJCV24	34.07M	26.79	0.951	36.61	0.971	31.45	0.885	29.22	0.884	22.59	0.831	29.33	0.904
InstructIR-5D [26]	ECCV24	15.84M	27.10	0.956	36.84	0.973	31.40	0.873	29.40	0.886	23.00	0.836	29.55	0.908
AdaIR [25]	ICLR25	28.77M	30.53	0.978	38.02	0.981	31.35	0.889	28.12	0.858	23.00	0.845	30.20	0.910
MoCE-IR-S	CVPR25	11.47M	31.33	0.978	37.21	0.978	31.25	0.884	28.90	0.877	21.68	0.851	30.08	0.913
MoCE-IR [29]	CVPR25	25.35M	30.48	0.974	38.04	0.982	31.34	0.887	30.05	0.899	23.00	0.852	30.58	0.919
Ours-B	Ours	7.42M	31.85	0.982	38.67	0.984	31.46	0.892	28.50	0.866	23.03	0.857	30.70	0.916

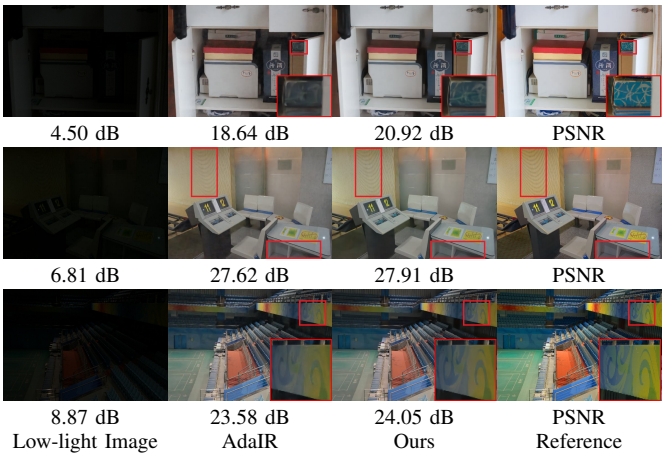


Fig. 7. Visual results on the real LOLv1 [97] dataset in the five-task setting.

Urban100, it achieves a 0.37 dB PSNR gain over AdaIR [25].

In addition, we evaluate our five-task model on tasks unseen during training, namely defocus deblurring, desnowing, and raindrop removal, using the DPPD [81], SRRS [63], and AGAN [73] datasets, respectively. As shown in Table XIII, our method outperforms two state-of-the-art all-in-one approaches across all three out-of-distribution tasks, highlighting its stronger generalization capability.

We further directly apply the five-task model to real-world images captured by UAVs [100]. Figure 8 illustrates that our approach demonstrates greater robustness than two strong competing methods in real haze removal. In particular, for the challenging heavy-haze scenes in the second example, it still produces visually reasonable results. Notably, it achieves

TABLE XII
PSNR RESULTS OF THE MODEL PRE-TRAINED IN THE FIVE-TASK ALL-IN-ONE SETTING, EVALUATED ON BSD68 [101], URBAN100 [98], AND KODAK24 [99] AT NOISE LEVELS OF 15, 25, AND 50.

Method	BSD68			Urban100			Kodak24		
	15	25	50	15	25	50	15	25	50
IDR [96]	34.11	31.60	28.14	33.82	31.29	28.07	34.78	32.42	29.13
InstructIR-5D [26]	34.00	31.40	28.15	33.77	31.40	28.13	34.70	32.26	29.16
MoCE-IR [29]	34.00	31.34	28.07	34.01	31.59	28.20	34.87	32.38	29.20
AdaIR [25]	34.01	31.35	28.06	34.10	31.68	28.29	34.89	32.38	29.21
Ours-B	34.13	31.46	28.20	34.35	31.97	28.66	35.05	32.57	29.41

TABLE XIII
GENERALIZATION EVALUATION OF OUR FIVE-TASK MODEL ON THREE UNSEEN TASKS.

Dataset	AdaIR [25]		MoCE-IR [29]		**Ours-B**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DPDD [81] (Defocus)	15.49	0.601	15.58	0.598	18.65	0.662
AGAN [73] (Raindrop)	23.08	0.827	22.39	0.815	23.10	0.827
SRRS [63] (Snow)	21.86	0.845	21.82	0.838	21.98	0.848

stronger generalization capability while requiring significantly fewer parameters (see Table XI).

4) Discussion: Following all-in-one methods [22], [24], [25], we compute t-SNE visualizations of intermediate features generated by our models under two all-in-one settings. As shown in Figure 9, the features form distinct clusters according to degradation types, indicating that our model can learn discriminative degradation information directly from corrupted inputs, despite the absence of explicit degradation-related priors. Instead, the model exhibits implicit dynamic learning capabilities: high-order interactions act as an adaptive

TABLE XIV
QUANTITATIVE RESULTS ON THE LOLBLUR [27] DATASET FOR COMPOSITE DEGRADATIONS.

Method	MIMO [8]	NAFNet [104]	LEDNet [27]	Restormer [14]	VQCNIR [30]	DarkIR-M [105]	DarkIR-L [105]	**Ours-T**	**Ours-B**
PSNR	22.41	25.36	25.74	26.72	27.79	27.00	27.30	27.20	28.49
SSIM	0.835	0.882	0.850	0.902	0.875	0.883	0.898	0.903	0.919
Params	6.8M	12.05M	7.4M	26.13M	47.9M	3.31M	12.96M	0.99M	7.42M

Hazy Image MoCE-IR AdaIR Ours-B

Fig. 8. Qualitative results on challenging real-world hazy images [100], with models applied in a zero-shot manner.

TABLE XV
AVERAGE LPIPS SCORES IN THE THREE-TASK ALL-IN-ONE SETTING.

Method	PromptIR [22]	AdaIR [25]	DA-CLIP [102]	DiffUIR [106]	UniLDiff [103]	**Ours-B**
LPIPS↓	0.086	0.083	0.066	0.096	0.065	0.074

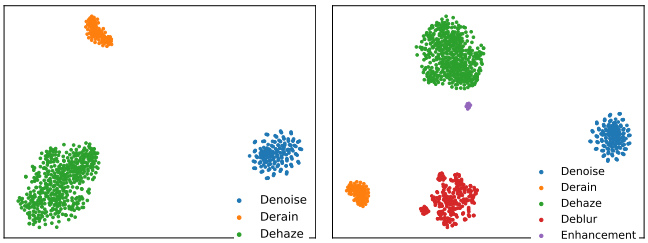
gating mechanism, attention weights in the similarity-aware weighting module function as a dynamic selection mechanism, and large-kernel convolutions in the encoding stage offer a broader receptive field for identifying degradation types. We believe that incorporating explicit priors may further enhance performance, positioning our model as a strong and efficient baseline for the all-in-one track.

Our architectural design is inspired by the human visual system with the aim of achieving high efficiency. Under a non-diffusion-based training paradigm, the proposed model demonstrates strong performance across a range of image restoration tasks. Given its biological motivation, we further assess its effectiveness using the perceptual metric.

Specifically, we compare our all-in-one model with competing approaches under the three-task all-in-one setting using LPIPS. As shown in Table XV, our method outperforms recent non-diffusion-based methods, including PromptIR [22] and AdaIR [25]. Although diffusion-based models generally excel in perceptual metrics, our model achieves competitive performance without incorporating text guidance or relying on large-scale pretrained models for information extraction, as employed in DA-CLIP [102] and UniLDiff [103].

C. Composite degradation image restoration

In this scenario, we evaluate VIVNet under two-degradation and three-degradation settings, where each image is simultaneously affected by two and three degradation types, respectively.



Three-task setting Five-task setting

Fig. 9. t-SNE visualization of intermediate features generated by our models in two all-in-one settings.

TABLE XVI
GENERALIZATION EVALUATION RESULTS ON THE UHD-LL [56] DATASET.

Method	PromptIR [22]	DiffIR [107]	DiffUIR [106]	AutoDIR [108]	FoundIR [109]	**Ours-T**
PSNR	12.16	12.69	11.18	16.10	16.23	18.29
SSIM	0.627	0.599	0.593	0.715	0.763	0.731

1) Two-degradation setting: Capturing images in under-lit environments often requires long exposure times, which can introduce blur artifacts. To address this composite degradation, we evaluate our model on the low-light deblurring task using the LOLBlur [27] dataset. For a fair and comprehensive comparison, we train two model variants. Table XIV shows that our tiny model surpasses DarkIR-M [105], a specially designed method, by 0.2 dB in PSNR while using only 30% of the parameters. Similarly, our base model outperforms DarkIR-L [105] by 1.19 dB with fewer parameters. Furthermore, our model exceeds VQCNIR [30] by 0.7 dB while using only 15% of its parameters. Figure 10 shows that our method removes composite degradations more effectively, producing sharper details from the blurry input and rendering colors closer to the reference image, particularly in the ground regions.

To evaluate the generalization capability of our model, we directly apply the model trained on the LOLBlur dataset [27] to the real-world UHD-LL dataset [56], which is affected by low-light conditions and noise. Quantitative comparisons with state-of-the-art universal image restoration methods are reported in Table XVI. Our model substantially outperforms the recent FoundIR method [109] in terms of PSNR, demonstrating its strong generalization ability. As illustrated in Figure 12, our model produces more visually faithful restoration results than competing approaches.

2) Three-degradation setting: We further evaluate our model in a challenging three-degradation scenario using the CDD-11 [28] dataset, which contains 11 degradation categories, with each image affected by up to three types of degradations simultaneously. Table XVII shows that our base model

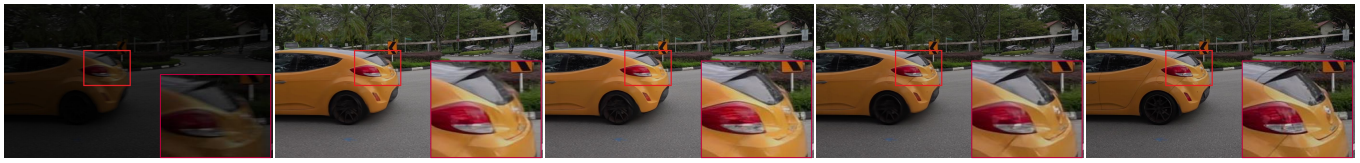


Fig. 10. Visual comparisons on the LOLBlur [27] dataset for image restoration under composite degradations.

Fig. 11. Visual comparisons on the CDD-11 [28] dataset for image restoration under composite degradations.

TABLE XVII
QUANTITATIVE RESULTS ON CDD-11 [28] FOR COMPOSITE DEGRADATION IMAGE RESTORATION. RESULTS ARE REPORTED IN PSNR AND SSIM.

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	
AirNet [21]	8.93M	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 [22]	35.6M	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
WGWSNet [110]	25.76M	24.39	.774	27.90	.982	33.15	.964	34.43	.973	24.27	.800	25.06	.772	24.60	.765	27.23	.955	27.65	.960	23.90	.772	23.97	.771	26.96	.863
WeatherDiff [111]	82.96M	23.58	.763	21.99	.904	24.85	.885	24.80	.888	21.83	.756	22.69	.730	22.12	.707	21.25	.868	21.99	.868	21.23	.716	21.04	.698	22.49	.799
OneRestore [28]	5.98M	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
MIRAGE [112]	6M	27.13	.830	32.39	.989	34.23	.969	35.57	.978	26.04	.823	26.21	.807	26.07	.799	29.49	.962	29.72	.967	25.17	.793	25.41	.793	28.86	.883
MoCE-IR-S [29]	11.47M	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
Our-T	0.99M	27.25	.829	33.00	.990	33.86	.966	35.36	.976	26.45	.824	26.46	.806	26.22	.797	30.62	.963	30.70	.969	25.55	.795	25.74	.796	29.20	.883
Our-B	7.42M	27.67	.837	37.42	.995	35.07	.972	36.79	.978	27.20	.834	26.82	.816	26.75	.810	32.86	.971	33.42	.976	26.37	.808	26.41	.809	30.62	.891

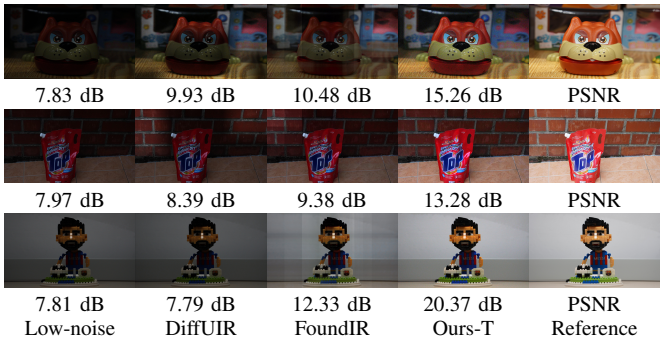


Fig. 12. Visual results on UHD-LL [56] for generalization evaluations.

achieves the highest scores on ten degradation categories. When averaged across all 11 categories, it delivers a 1.57 dB PSNR improvement over MoCE-IR-S [29] while using fewer parameters. Notably, our tiny model still surpasses this strong competitor on most degradation combinations with only 8% of the parameters. These results clearly demonstrate the high efficiency and strong effectiveness of our approach. Figure 11 illustrates that the image restored by our method achieves a higher quality score and exhibits fewer visible degradations.

D. Ultra-high-definition image restoration

UHD images with high pixel density and resolution (*e.g.*, 4K or higher) have become increasingly common in everyday applications [55]. Following the recent specialized UHD image restoration method ERR [61], we evaluate our model on four tasks: dehazing, deraining, deblurring, and low-light image enhancement. The results are presented in Tables XVIII, XIX, XX, and XXI, respectively. ERR employs frequency processing and Kolmogorov–Arnold Networks (KAN) [113] within a heterogeneous U-shaped architecture to enhance performance and efficiency. In contrast, our model adopts a homogeneous design across all blocks, resulting in a simpler architecture, and surpasses ERR on all four datasets in both PSNR and SSIM while using fewer parameters. Notably, our method achieves gains of 0.76 dB and 0.80 dB in PSNR over ERR on the UHD-Haze and UHD-Blur datasets, respectively. Furthermore, it outperforms specialized Transformer-based [57], [60] and Mamba-based [46] methods while maintaining high parameter efficiency, demonstrating strong adaptability to UHD tasks.

E. Domain-specific image restoration

Beyond natural images, image restoration plays a crucial role in specialized domains such as medical imaging, remote sensing, and underwater image enhancement, where we further evaluate the proposed model.

1) Medical image restoration: We conduct comprehensive experiments on three medical image restoration tasks: PET image synthesis, CT image denoising, and MRI image super-resolution. Following [115], we train separate models for each task. The comparison results are shown in Tables XXII,

TABLE XVIII
COMPARISONS ON UHD-HAZE [57] FOR UHD IMAGE DEHAZING.

Method	PSNR	SSIM	Params
Uformer [15]	19.83	0.737	20.6M
DehazeFormer-B [44]	15.37	0.725	2.5M
UHDFormer [57]	22.59	0.943	0.339M
UHDDIP [59]	24.69	0.952	0.81M
ERR [61]	25.12	0.950	1.131M
Ours-T	25.88	0.959	0.99M

TABLE XIX
COMPARISONS ON 4K-RAIN13K [58] FOR UHD IMAGE DERAINING.

Method	PSNR	SSIM	Params
Restormer [14]	33.02	0.933	26.12M
DRSformer [69]	32.94	0.933	33.65M
UDR-S2Former [114]	33.36	0.946	8.53M
UDR-Mixer [58]	34.28	0.951	4.90M
ERR [61]	34.48	0.952	1.131M
Ours-T	34.78	0.956	0.99M

TABLE XX
COMPARISONS ON UHD-BLUR [57] FOR UHD IMAGE DEBLURRING.

Method	PSNR	SSIM	Params
Stripformer [43]	25.05	0.725	19.7M
FFTformer [42]	25.41	0.725	16.6M
UHDFormer [57]	28.82	0.844	0.339M
UHDDIP [59]	29.51	0.858	0.81M
ERR [61]	29.72	0.861	1.131M
Ours-T	30.52	0.877	0.99M

TABLE XXI
RESULTS ON UHD-LL [56] FOR UHD LOW-LIGHT IMAGE ENHANCEMENT.

Method	PSNR	SSIM	Params
LLFormer [60]	22.79	0.853	13.15M
UHDFour [56]	26.22	0.900	17.54M
Wave-Mamba [46]	27.35	0.913	1.258M
UHDFormer [57]	27.11	0.927	0.339M
ERR [61]	27.57	0.932	1.131M
Ours-T	27.74	0.932	0.99M

TABLE XXII
PET IMAGE SYNTHESIS RESULTS ON THE POLARSTAR M660 [115] DATASET.

Method	PSNR↑	SSIM↑	RMSE↓	Params
CycleWGAN [116]	36.62	0.929	0.091	1.00M
DCITN [117]	36.09	0.929	0.097	0.08M
DRMC [118]	36.00	0.935	0.100	0.62M
ARGAN [119]	36.73	0.941	0.090	31.14M
R-RWKV-l [115]	36.96	0.943	0.089	1.16M
Ours-T	37.19	0.947	0.087	0.99M

TABLE XXIII
CT IMAGE DENOISING RESULTS ON THE AAPM [120] DATASET.

Method	PSNR↑	SSIM↑	RMSE↓	Params
TransCT [121]	32.62	0.908	9.533	13.23M
Eformer [122]	33.35	0.918	8.803	0.34M
CTformer [123]	33.25	0.913	8.897	1.45M
DenoMamba [124]	33.53	0.915	8.612	112.62M
R-RWKV-l [115]	33.64	0.918	8.514	1.16M
Ours-T	33.76	0.919	8.400	0.99M

TABLE XXIV
MRI IMAGE SUPER-RESOLUTION RESULTS ON THE IXI [125] DATASET.

Method	PSNR↑	SSIM↑	RMSE↓	Params
FAWDN [126]	30.04	0.914	35.956	5.07M
SwinMR [127]	30.93	0.925	32.734	11.40M
SDAUT [128]	30.96	0.926	32.593	67.23M
F-UNet [129]	31.26	0.931	31.568	32.12M
R-RWKV-l [115]	31.36	0.931	31.256	1.16M
Outs-T	31.39	0.932	31.118	0.99M

TABLE XXV
REMOTE SENSING IMAGE DEHAZING RESULTS ON THE THREE HAZE LEVELS OF SATEHAZE1K [130].

Method	Thin		Moderate		Thick	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EMPF [131]	22.69	0.896	25.17	0.932	20.23	0.822
Trinity [132]	22.65	0.896	24.73	0.934	20.57	0.824
FocalNet [12]	24.16	0.916	25.99	0.947	21.69	0.847
FMambaIR [133]	24.58	0.912	25.83	0.939	22.65	0.850
Ours-S	25.45	0.928	26.86	0.934	23.05	0.866

TABLE XXVI
RESULTS ON LSUI [134] FOR UNDERWATER IMAGE ENHANCEMENT.

Method	PSNR	SSIM	Params
UColor [135]	22.91	0.891	157.4M
UShape [134]	24.16	0.932	65.6M
DM-Water [136]	27.65	0.887	10.13M
WF-Diff [137]	27.26	0.944	43.72M
SS-UIE [138]	28.87	0.952	4.25M
Ours-S	29.09	0.906	2.86M

XXIII, and XXIV, respectively. R-RWKV-l [115] introduces the Receptance Weighted Key Value (RWKV) [139] architecture to this domain for the first time. Unlike this method, our approach is built on a brain-inspired mechanism and achieves superior performance across all three tasks in terms of PSNR, SSIM, and RMSE, while using slightly fewer parameters. Specifically, on the PolarStar M660 dataset [115] for PET image synthesis, our model achieves 37.19 dB in PSNR, outperforming R-RWKV-l by 0.23 dB with 15% fewer parameters. On CT image denoising, it surpasses the Mamba-based DenoMamba [124] while requiring only 0.9% of the parameters. These results highlight the strong adaptability of our model to medical image restoration tasks.

2) Remote sensing image restoration: Remote sensing images are often degraded by haze, which can hinder the identification of ground targets. We evaluate our model for remote sensing image dehazing using the Satehaze1k [130] dataset, which contains three haze levels: thin, moderate, and thick. As shown in Table XXV, our method consistently outperforms both specialized and general approaches across most metrics. In particular, it achieves a 0.87 dB PSNR gain over the second-best methods at both the thin and moderate haze levels. Even under the thick haze condition, our model surpasses the Mamba-based FMambaIR [133] framework by 0.4 dB in PSNR and 0.016 in SSIM.

3) Underwater image enhancement: Underwater image enhancement plays an important role in ocean exploration by enhancing visibility and image quality in challenging aquatic environments. To evaluate the adaptability of our model to this task, we use the widely adopted LSUI [134] dataset and compare its performance with that of specialized algorithms. As shown in Table XXVI, our model achieves 29.09 dB PSNR, surpassing the recent dual-domain SS-UIE [138] by 0.22 dB while maintaining higher parameter efficiency. Compared to diffusion-based models such as WF-Diff [137] and DM-Water [136], the efficiency advantage becomes even more significant, further demonstrating the effectiveness of our approach for underwater image enhancement.

F. Inference time analysis

In practical deployments, inference time is a critical factor that must be carefully considered. To this end, we assess the runtime efficiency of the proposed method on several widely used benchmarks, covering general-purpose, all-in-one, and composite degradation image restoration tasks, and compare its performance with that of leading competitors. The results are summarized in Table XXVII.

Our model outperforms multiple variants of the recent Transformer-based MT V2 [66] in image dehazing and desnowing, achieving higher accuracy while also delivering significantly faster inference speed. On the SRRS [63] dataset

TABLE XXVII
INFERENCE TIME COMPARISON BETWEEN STATE-OF-THE-ART ALGORITHMS AND OUR METHOD ACROSS VARIOUS IMAGE RESTORATION TASKS. THE INFERENCE TIME IS MEASURED ON AN NVIDIA TESLA A100 (40 GB) GPU. THE VALUES INDICATE THE ACCURACY (+) AND SPEED (×) GAINS OVER THE COMPETITOR.

Setting	Task/Dataset	Method	PSNR (dB)	Time (s)
General	Dehazing	MT-B V2 [66]	41.00	1.230
	SOTS-Indoor [94]	Ours-S	42.03 +1.03	0.089 ×13.82
	Desnowing	MT-L V2 [66]	32.55	2.244
	SRRS [63]	Ours-B	33.05 +0.50	0.145 ×15.48
All-in-one	Five-task	MoCE-IR [29]	38.04	0.225
	Rain100L [140]	Ours-B	38.67 +0.63	0.085 ×2.65
Composite	Two-degradation	VQCNIR [30]	27.79	0.363
	LOLBlur [27]	Ours-B	28.49 +0.70	0.322 ×1.13
	Three-degradation	MoCE-IR-S [29]	32.66	0.893
	CDD-11 (Haze) [28]	Ours-T	33.00 +0.34	0.116 ×7.70

TABLE XXVIII
ABLATION STUDY ON THE PROPOSED COMPONENTS. SW AND HP REFER TO THE SIMILARITY-AWARE WEIGHTING MECHANISM AND HIGH-ORDER PROCESSING MODULE, RESPECTIVELY.

Net	SW	HP	PSNR	FLOPs	Params
(a)			31.40 +0.00	13.32G	0.86M
(b)	✓		34.77 +3.37	13.35G	0.98M
(c)		✓	34.41 +3.01	13.08G	0.83M
(d)	✓	✓	35.64 +4.24	13.10G	0.94M

TABLE XXIX
ABLATION STUDY ON THE ORDER OF INTERACTIONS BETWEEN VISION SIGNALS. NET (F) INCREASES THE NUMBER OF BASIC BLOCKS IN NET (A) TO MATCH THE PARAMETER COUNT OF NET (D).

Net	Order	PSNR	FLOPs	Params
(a)	1	31.53 +0.00	12.65G	0.81M
(b)	2	32.89 +1.36	12.79G	0.82M
(c)	3	33.74 +2.21	12.93G	0.82M
(d)	4	34.41 +2.88	13.08G	0.83M
(e)	5	34.62 +3.09	13.22G	0.83M
(f)	1	32.07 +0.54	13.61G	0.83M

TABLE XXX
ABLATION STUDY ON THE KERNEL SIZE OF A DEPTH-WISE CONVOLUTION IN THE ENCODING STAGE, WITH THE OTHER KEPT FIXED AT 3×3. NET (G) ADOPTS A LARGER CHANNEL DIMENSION THAN NET (B).

Net	Kernel	PSNR	FLOPs	Params
(a)	1×1	35.67 +0.00	12.97G	0.94M
(b)	3×3	35.64 −0.03	13.10G	0.94M
(c)	5×5	35.95 +0.28	13.35G	0.95M
(d)	7×7	36.16 +0.49	13.73G	0.97M
(e)	9×9	36.63 +0.96	14.18G	0.99M
(f)	11×11	36.54 +0.87	14.86G	1.01M
(g)	3×3	35.74 +0.07	14.62G	1.06M

for image desnowing, the base version of our model runs 15.48× faster than MT-L V2. In the five-task all-in-one setting, it is 2.65× faster than the state-of-the-art MoCE-IR [29], despite the latter employing a mixture-of-experts framework to enhance processing speed. The runtime advantage also extends to composite degradation scenarios. On the CDD-11 [28] dataset, our tiny model achieves 7.70× faster inference than MoCE-IR-S [29] while surpassing it by 0.34 dB in PSNR in the dehazing category. These results highlight the strong suitability of our model for real-world deployment, where both accuracy and efficiency are critical.

G. Ablation studies

In this section, we conduct ablation experiments to verify the effectiveness of the proposed components and investigate alternative design strategies. Unless otherwise stated, the tiny version of our model is trained on the dehazing task using RESIDE-Indoor [94] for 100K iterations and evaluated on SOTS-Indoor [94]. All other settings are kept consistent with those of the dehazing model in the general-purpose setting. SW and HP denote the similarity-aware weighting mechanism and the high-order processing module, respectively. The orange row marks the final parameter configuration adopted in each ablation study. Values marked with + and − indicate performance improvements and degradations, respectively.

Effects of individual components. We evaluate the impact of the proposed components in Table XXVIII. The baseline (Net (a)), obtained by removing both SW and HP from our model and setting the kernel size of the two depth-wise convolutions to 3×3, achieves 31.40 dB in PSNR. Adding SW improves performance by 3.37 dB with negligible computational overhead, while incorporating HP yields a 3.01 dB gain and reduces both FLOPs and parameter count. Combining both components produces the complete model, which achieves the highest performance, surpassing the baseline by a large margin while maintaining lower computational complexity. These results demonstrate the effectiveness of our design in improving performance without sacrificing efficiency.

The number of interaction orders in HP. In HP, high-order processing is implemented by combining element-wise multiplication with depth-wise convolution. Table XXIX analyzes the impact of varying the order number. As the order increases, PSNR, FLOPs, and parameter count all rise. When the order reaches 5, performance gains slow considerably. Therefore, we set the order to 4 in our final model to balance accuracy and computational complexity.

To further validate the effectiveness of HP, we perform an additional experiment by increasing the number of basic blocks in Table XXIX(a) from [1, 1, 1, 1, 1, 5] to [1, 1, 1, 1, 1, 6]. This variant (Net (f)) matches the parameter count of our final model (Net (d)) but incurs higher FLOPs, achieving only 32.07 dB PSNR, significantly lower than our model, thereby confirming the advantage of our high-order interaction.

Furthermore, we visualize the intermediate features of the high-order interaction mechanism in Figure 13. As the order increases, the feature responses become progressively stronger, and informative signals such as salient object edges are more prominently highlighted and appear sharper.

Kernel size analysis of depth-wise convolution in the encoding stage. In the encoding stage, depth-wise convolutions with different kernel sizes are employed to capture visual information at multiple spatial scales. We investigate the impact of the kernel size of one depth-wise convolution while keeping the other fixed at 3×3, with results reported in Table XXX.

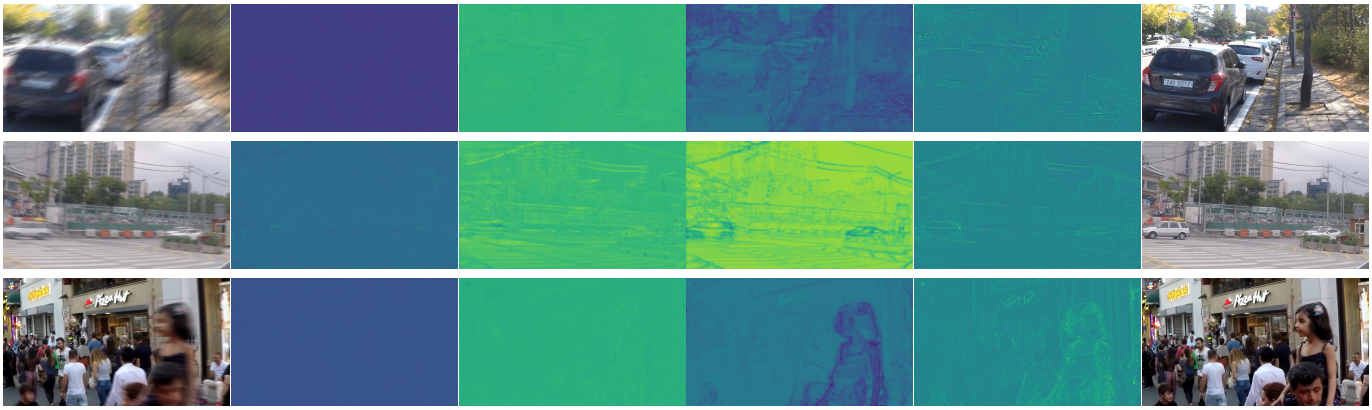


Fig. 13. Visualizations of intermediate features from the high-order processing mechanism. From left to right: the blurry input image, intermediate features with order numbers from 1 to 4, and the corresponding sharp image. As the order increases, feature responses become progressively stronger, and informative signals, such as the edges of salient objects, are more prominently highlighted and appear sharper.

TABLE XXXI
COMPARISON BETWEEN SW AND A SIMPLIFIED CHANNEL ATTENTION [104] UNDER VARIOUS BASELINE SETTINGS.

Baseline Setting	Method	PSNR	FLOPs	Params
w/o HP (Table XXVIII(a))	SCA	34.46	13.39G	1.05M
	Ours	34.77	13.35G	0.98M
w/ HP (Table XXIX(a))	SCA	34.66	12.71G	1.00M
	Ours	35.01	12.68G	0.93M

TABLE XXXII
COMPARISON OF ALTERNATIVE SIMILARITY METRICS IN SW.

Method	Euclidean distance	Pearson correlation coefficient	Ours
PSNR	34.19	35.28	35.64

TABLE XXXIII
COMPARISON OF ALTERNATIVE APPROACHES TO THE HP MODULE.

Method	Addition	Concat+CNN	Cross Attention	HP	Final
PSNR	34.17	31.32	35.52	34.41	36.63
FLOPs	13.08G	14.01G	15.76G	13.08G	14.18G
Params	0.83M	0.91M	1.02M	0.83M	0.99M

Overall, PSNR improves as the kernel size increases. However, using a 11×11 kernel results in a performance drop despite higher computational cost. This degradation is likely due to the increased difficulty in fusing visual cues from the two depth-wise convolutions, caused by the large disparity in their receptive fields. Consequently, we adopt a 9×9 kernel in our model to balance performance and computational complexity.

In addition, we conduct an experiment by increasing the channel dimension of Net (b) to obtain Net (g), such that its computational overhead matches that of our choice, Net (e). Despite the increased capacity, this variant achieves markedly inferior performance compared to our model, while incurring a higher computational cost, further demonstrating the effectiveness of our design.

Alternative to SW. SW is designed to emphasize informative signals prior to high-order interaction. Unlike regular channel-wise attention, which generates weights by applying linear layers to the pooled features, our method derives the attention weight for each channel from its similarity to all channels, thereby better capturing its importance within the features. To empirically validate the effectiveness of our design, we compare it with an alternative approach, the simplified channel attention (SCA) used in [104], under various baseline settings. As shown in Table XXXI, SW consistently outperforms the alternative in both baseline settings while reducing approximately 0.03G FLOPs and 0.07M parameters.

In addition, we perform comparative experiments by substituting cosine similarity with Euclidean distance and Pearson correlation coefficient. As shown in Table XXXII, cosine similarity consistently yields superior performance, as it models relative feature orientation and is more robust to scale variations as well as distance concentration in high-dimensional deep representations.

Alternatives to HP. We implement the high-order interaction using a simple combination of depth-wise convolutions and element-wise multiplication. To assess its effectiveness, we replace it with several alternatives and report the results in Table XXXIII. Substituting the multiplication with addition yields 34.17 dB in PSNR, which is 0.24 dB lower than our design. Replacing our HP entirely with concatenation followed by a 1×1 convolution significantly degrades performance, despite incurring higher complexity. We also explore using channel-wise cross-attention to achieve interactions. While this variant achieves better performance than our design, it introduces substantially higher computational overhead, which conflicts with our goal of developing efficient yet effective baselines. Notably, our final model achieves a better trade-off than the cross-attention variant, surpassing it in computational efficiency while maintaining superior performance. These results clearly demonstrate the effectiveness of our HP module as well as the overall design.

Effects of the reduction factor in SW. In Eq. (9), we employ linear layers along with a reduction factor to control the weight generation process and determine the size of the intermediate layer. Table XXXIV shows that increasing this coefficient raises computational overhead. Accuracy improves

TABLE XXXIV
ABLATION STUDY ON THE REDUCTION FACTOR IN SW (EQ. (9)).

Coefficient	PSNR	FLOPs	Params
0.2	34.71 +0.00	12.66G	0.85M
0.4	34.93 +0.22	12.67G	0.89M
0.6	35.01 +0.30	12.68G	0.93M
0.8	34.79 +0.08	12.68G	0.97M

TABLE XXXV
ABLATION STUDIES ON DIFFERENT NUMBERS OF SCALES.

Method	PSNR	SSIM	FLOPs	Params
Three-scale	39.27	0.996	25.97G	2.86M
Four-scale	39.43	0.996	26.29G	3.95M

with increasing factors, peaking at 0.6, but declines beyond this point; for example, a factor of 0.8 results in a performance drop. This degradation is likely due to significant information loss introduced by the second linear layer. Therefore, maintaining a balanced channel reduction between the two linear layers is more effective for generating accurate attention weights.

Three-scale or four-scale architecture? Following previous general and multi-task image restoration models [14], [15], [22], [25], [29], [62], [66], [104], [109], [141], we adopt a U-shaped architecture to enable efficient multi-scale learning. Existing approaches typically employ either a three-scale or four-scale pipeline. The key difference between these two variants is that the additional deepest layer in the four-scale design introduces a substantial number of parameters. To investigate this design choice, we conduct experiments using our small model and its four-scale counterpart, with the number of blocks set to [3, 3, 1, 1, 1, 1, 3, 7], ensuring comparable computational overhead between the two models. Both models are trained for 100K iterations, with all other settings identical to those of our final dehazing model on the SOTS-Indoor [94] dataset. As shown in Table XXXV, the three-scale version achieves comparable performance while reducing the parameter count by 28%. Based on these results, we adopt a three-scale architecture to achieve a better balance between performance and computational efficiency.

H. Failure cases

Although our model performs favorably against state-of-the-art methods across a wide range of image restoration tasks, certain failure cases remain. As shown in the third example of Figure 5, the rainy image contains extremely dense and fine-grained rain streaks with complex orientations and overlapping layers. These rain patterns exhibit strong similarity to the underlying textures, making it difficult for the model to accurately disentangle rain artifacts from structural details. As a result, our method fails to fully remove the rain streaks and yields an overly smoothed restoration. Furthermore, due to the scarcity of real-world paired training data, the model is unable to completely remove severe haze degradations in real-world scenes after training, as illustrated in Figure 14. Future work may focus on collecting additional real-world datasets, especially those captured under such challenging conditions, to further enhance restoration performance.

Fig. 14. Visual results on Dense-Haze [88] for real-world image dehazing.

V. CONCLUSION

In this study, we present a universal and efficient network for image restoration. Instead of relying on increasingly complex architectures such as Transformers or Mamba variants, we take inspiration from the human visual system. To this end, we analyze its perceptual processes and design a brain-inspired mechanism based on lightweight operations for high efficiency. Specifically, our model leverages depth-wise convolutions with varying kernel sizes to capture visual cues at multiple spatial scales, followed by a similarity-aware weighting mechanism that reduces redundancy and emphasizes informative signals. Finally, high-order interactions, realized through the iterative application of element-wise multiplication and depth-wise convolution, are incorporated to enhance semantic understanding and further highlight and sharpen salient objects.

We extensively evaluate the proposed model across diverse image restoration settings to demonstrate its effectiveness and generality. Specifically, we conduct experiments on key benchmarks covering general-purpose, all-in-one, and composite degradation image restoration tasks. We further assess the model's adaptability to UHD image restoration and extend the evaluation to domain-specific tasks, including medical imaging, remote sensing, and underwater image enhancement. Experimental results show that our model consistently achieves state-of-the-art performance while maintaining high efficiency and fast inference speed.

VI. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China (Grant No. U24B20175, No. 62322216, No. 62136001), Shenzhen Science and Technology Program (Grant No. JCYJ20241202125904007, No. RCYX20221008092849068), the Open Fund of the Key Laboratory of the Ministry of Education on Artificial Intelligence in Equipment (Grant No. 2024-AAIE-KF04-01), the Open Research Fund of the State Key Laboratory of Blockchain and Data Security, Zhejiang University, and Beijing Major Science and Technology Project (Grant No. Z251100008125009).

REFERENCES

[1] K. Zhang, W. Ren, W. Luo, W.-S. Lai, B. Stenger, M.-H. Yang, and H. Li, "Deep image deblurring: A survey," *IJCV*, 2022.
[2] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Jia, "Ffa-net: Feature fusion attention network for single image dehazing," in *AAAI*, 2020.
[3] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE TIP*, 2016.
[4] X. Liu, Y. Ma, Z. Shi, and J. Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *ICCV*, 2019.

[5] X. Fu, J. Huang, D. Zeng, Y. Huang, X. Ding, and J. Paisley, "Removing rain from single images via a deep detail network," in *CVPR*, 2017.

[6] W. Yang, R. T. Tan, S. Wang, Y. Fang, and J. Liu, "Single image deraining: From model-based to data-driven and beyond," *IEEE TPAMI*, 2020.

[7] R. Quan, X. Yu, Y. Liang, and Y. Yang, "Removing raindrops and rain streaks in one go," in *CVPR*, 2021.

[8] 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 *ICCV*, 2021.

[9] J. Lee, H. Son, J. Rim, S. Cho, and S. Lee, "Iterative filter adaptive network for single image defocus deblurring," in *CVPR*, 2021.

[10] H. Son, J. Lee, S. Cho, and S. Lee, "Single image defocus deblurring using kernel-sharing parallel atrous convolutions," in *ICCV*, 2021.

[11] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao, "Multi-stage progressive image restoration," in *CVPR*, 2021.

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

[13] Y. Cui, W. Ren, S. Yang, X. Cao, and A. Knoll, "Irnext: Rethinking convolutional network design for image restoration," in *ICML*, 2023.

[14] 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 *CVPR*, 2022.

[15] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *CVPR*, 2022.

[16] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *ICCVW*, 2021.

[17] Y. Li, Y. Fan, X. Xiang, D. Demandolx, R. Ranjan, R. Timofte, and L. Van Gool, "Efficient and explicit modelling of image hierarchies for image restoration," in *CVPR*, 2023.

[18] 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.

[19] B. Li, H. Zhao, W. Wang, P. Hu, Y. Gou, and X. Peng, "Mair: A locality-and continuity-preserving mamba for image restoration," in *CVPR*, 2025.

[20] Y. Gu, Y. Meng, J. Ji, and X. Sun, "Acl: Activating capability of linear attention for image restoration," in *CVPR*, 2025.

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

[22] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. Shahbaz Khan, "Promptir: Prompting for all-in-one image restoration," in *NeurIPS*, 2023.

[23] 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 TIP*, 2025.

[24] Y. Ai, H. Huang, X. Zhou, J. Wang, and R. He, "Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration," in *CVPR*, 2024.

[25] 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 *ICLR*, 2025.

[26] M. V. Conde, G. Geigle, and R. Timofte, "High-quality image restoration following human instructions," in *ECCV*, 2024.

[27] S. Zhou, C. Li, and C. Change Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *ECCV*, 2022.

[28] Y. Guo, Y. Gao, Y. Lu, H. Zhu, R. W. Liu, and S. He, "Onerestore: A universal restoration framework for composite degradation," in *ECCV*, 2024.

[29] 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 *CVPR*, 2025.

[30] 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.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[32] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proceedings of the national academy of sciences*, 2007.

[33] E. E. Stewart, M. Valsecchi, and A. C. Schütz, "A review of interactions between peripheral and foveal vision," *Journal of vision*, 2020.

[34] Y. Wang, Q. Zhao, M. Ma, and J. Xu, "Olfactory perception prediction model inspired by olfactory lateral inhibition and deep feature combination," *Applied Intelligence*, 2023.

[35] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron*, 2012.

[36] X. Ma, X. Dai, Y. Bai, Y. Wang, and Y. Fu, "Rewrite the stars," in *CVPR*, 2024.

[37] X. Ma, X. Dai, J. Yang, B. Xiao, Y. Chen, Y. Fu, and L. Yuan, "Efficient modulation for vision networks," in *ICLR*, 2024.

[38] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[39] L. Zhu, C.-W. Fu, D. Lischinski, and P.-A. Heng, "Joint bi-layer optimization for single-image rain streak removal," in *ICCV*, 2017.

[40] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in *CVPR*, 2022.

[41] C.-L. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3d position embedding," in *CVPR*, 2022.

[42] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *CVPR*, 2023.

[43] F.-J. Tsai, Y.-T. Peng, Y.-Y. Lin, C.-C. Tsai, and C.-W. Lin, "Stripformer: Strip transformer for fast image deblurring," in *ECCV*, 2022.

[44] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE TIP*, 2023.

[45] X. Chen, Z. Li, Y. Pu, Y. Liu, J. Zhou, Y. Qiao, and C. Dong, "A comparative study of image restoration networks for general backbone network design," in *ECCV*, 2024.

[46] W. Zou, H. Gao, W. Yang, and T. Liu, "Wave-mamba: Wavelet state space model for ultra-high-definition low-light image enhancement," in *ACM MM*, 2024.

[47] Z. Zou, H. Yu, J. Huang, and F. Zhao, "Freqmamba: Viewing mamba from a frequency perspective for image deraining," in *ACM MM*, 2024.

[48] Y. He, L. Peng, Q. Yi, C. Wu, and L. Wang, "Multi-scale representation learning for image restoration with state-space model," *arXiv preprint arXiv:2408.10145*, 2024.

[49] 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," in *NeurIPS*, 2024.

[50] Y.-C. Lin, Y.-S. Xu, H.-W. Chen, H.-K. Kuo, and C.-Y. Lee, "Eamamba: Efficient all-around vision state space model for image restoration," in *ICCV*, 2025.

[51] X. Chen, J. Pan, J. Dong, J. Yang, and J. Tang, "Foundir-v2: Optimizing pre-training data mixtures for image restoration foundation model," *arXiv preprint arXiv:2512.09282*, 2025.

[52] G. Wu, J. Jiang, K. Jiang, and X. Liu, "Harmony in diversity: Improving all-in-one image restoration via multi-task collaboration," in *ACM MM*, 2024.

[53] Y. Ai, H. Huang, and R. He, "Lora-ir: taming low-rank experts for efficient all-in-one image restoration," *arXiv preprint arXiv:2410.15385*, 2024.

[54] Y. Tian, J. Han, H. Chen, Y. Xi, N. Ding, J. Hu, C. Xu, and Y. Wang, "Instruct-ipt: All-in-one image processing transformer via weight modulation," *arXiv preprint arXiv:2407.00676*, 2024.

[55] L. Wang, W. Zhou, C. Wang, K.-M. Lam, Z. Su, and J. Pan, "Deep learning-driven ultra-high-definition image restoration: A survey," *arXiv preprint arXiv:2505.16161*, 2025.

[56] C. Li, C.-L. Guo, man zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, "Embedding fourier for ultra-high-definition low-light image enhancement," in *ICLR*, 2023.

[57] 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.

[58] 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," *arXiv preprint arXiv:2405.17074*, 2024.

[59] L. Wang, C. Wang, J. Pan, W. Zhou, X. Sun, W. Wang, and Z. Su, "Ultra-high-definition restoration: New benchmarks and a dual interaction prior-driven solution," *arXiv e-prints*, 2024.

[60] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," in *AAAI*, 2023.

[61] 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.

[62] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Revitalizing convolutional network for image restoration," *IEEE TPAMI*, 2024.

[63] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *ECCV*, 2020.

[64] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "Desnownet: Context-aware deep network for snow removal," *IEEE TIP*, 2018.

[65] 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 *ICCV*, 2023.

[66] 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," *arXiv preprint arXiv:2501.04486*, 2025.

[67] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *CVPR*, 2018.

[68] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, and R. W. Lau, "Spatial attentive single-image deraining with a high quality real rain dataset," in *CVPR*, 2019.

[69] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *CVPR*, 2023.

[70] X. Chen, J. Pan, and J. Dong, "Bidirectional multi-scale implicit neural representations for image deraining," in *CVPR*, 2024.

[71] N. Gao, X. Jiang, X. Zhang, and Y. Deng, "Efficient frequency-domain image deraining with contrastive regularization," in *ECCV*, 2024.

[72] D. Li, Y. Liu, X. Fu, S. Xu, and Z.-J. Zha, "Fouriermamba: Fourier learning integration with state space models for image deraining," *arXiv preprint arXiv:2405.19450*, 2024.

[73] 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.

[74] J. Xiao, X. Fu, A. Liu, F. Wu, and Z.-J. Zha, "Image de-raining transformer," *IEEE TPAMI*, 2022.

[75] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxim: Multi-axis mlp for image processing," in *CVPR*, 2022.

[76] 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.

[77] 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.

[78] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE TIP*, 2021.

[79] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "Snr-aware low-light image enhancement," in *CVPR*, 2022.

[80] 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 *ICCV*, 2023.

[81] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *ECCV*, 2020.

[82] Y. Quan, Z. Wu, and H. Ji, "Neumann network with recursive kernels for single image defocus deblurring," in *CVPR*, 2023.

[83] X. Jiang, X. Zhang, N. Gao, and Y. Deng, "When fast fourier transform meets transformer for image restoration," in *ECCV*, 2024.

[84] H. Feng, H. Zhou, T. Ye, S. Chen, and L. Zhu, "Residual diffusion deblurring model for single image defocus deblurring," in *AAAI*, 2025.

[85] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, 2017.

[86] Y. Cui, Y. Tao, Z. Bing, W. Ren, X. Gao, X. Cao, K. Huang, and A. Knoll, "Selective frequency network for image restoration," in *ICLR*, 2023.

[87] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain adaptation for image dehazing," in *CVPR*, 2020.

[88] 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 *ICIP*, 2019.

[89] C. O. Ancuti, C. Ancuti, and R. Timofte, "Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in *CVPRW*, 2020.

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

[91] Y. Liu, L. Zhu, S. Pei, H. Fu, J. Qin, Q. Zhang, L. Wan, and W. Feng, "From synthetic to real: Image dehazing collaborating with unlabeled real data," in *ACM MM*, 2021.

[92] T. Ye, Y. Zhang, M. Jiang, L. Chen, Y. Liu, S. Chen, and E. Chen, "Perceiving and modeling density for image dehazing," in *ECCV*, 2022.

[93] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Image restoration via frequency selection," *IEEE TPAMI*, 2023.

[94] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE TIP*, 2018.

[95] 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," *IJCV*, 2024.

[96] J. Zhang, J. Huang, M. Yao, Z. Yang, H. Yu, M. Zhou, and F. Zhao, "Ingredient-oriented multi-degradation learning for image restoration," in *CVPR*, 2023.

[97] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *BMVC*, 2018.

[98] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015.

[99] F. Rich, "Kodak lossless true color image suite," http://r0k.us/graphics/kodak, 1999.

[100] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *ECCV*, 2018.

[101] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001.

[102] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Controlling vision-language models for universal image restoration," in *ICLR*, 2024.

[103] Z. Cheng, L. Zhou, D. Chen, N. Tang, X. Luo, and Y. Qu, "Unildiff: Unlocking the power of diffusion priors for all-in-one image restoration," *arXiv preprint arXiv:2507.23685*, 2025.

[104] L. Chen, X. Chu, X. Zhang, and J. Sun, "Simple baselines for image restoration," in *ECCV*, 2022.

[105] D. Feijoo, J. C. Benito, A. Garcia, and M. V. Conde, "Darkir: Robust low-light image restoration," in *CVPR*, 2025.

[106] D. Zheng, X.-M. Wu, S. Yang, J. Zhang, J.-F. Hu, and W.-S. Zheng, "Selective hourglass mapping for universal image restoration based on diffusion model," in *CVPR*, 2024.

[107] B. Xia, Y. Zhang, S. Wang, Y. Wang, X. Wu, Y. Tian, W. Yang, and L. Van Gool, "Diffir: Efficient diffusion model for image restoration," in *ICCV*, 2023.

[108] Y. Jiang, Z. Zhang, T. Xue, and J. Gu, "Autodir: Automatic all-in-one image restoration with latent diffusion," in *ECCV*, 2024.

[109] H. Li, X. Chen, J. Dong, J. Tang, and J. Pan, "Foundir: Unleashing million-scale training data to advance foundation models for image restoration," in *ICCV*, 2025.

[110] Y. Zhu, T. Wang, X. Fu, X. Yang, X. Guo, J. Dai, Y. Qiao, and X. Hu, "Learning weather-general and weather-specific features for image restoration under multiple adverse weather conditions," in *CVPR*, 2023.

[111] O. Özdenizci and R. Legenstein, "Restoring vision in adverse weather conditions with patch-based denoising diffusion models," *IEEE TPAMI*, 2023.

[112] B. Ren, Y. Li, X. Zheng, Y. Fu, D. P. Paudel, M.-H. Yang, L. Van Gool, and N. Sebe, "Manifold-aware representation learning for degradation-agnostic image restoration," *arXiv preprint arXiv:2505.18679*, 2025.

[113] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[114] 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.

[115] Z. Yang, J. Li, H. Zhang, D. Zhao, B. Wei, and Y. Xu, "Restore-rwkv: Efficient and effective medical image restoration with rwkv," *arXiv preprint arXiv:2407.11087*, 2025.

[116] L. Zhou, J. D. Schaefferkoetter, I. W. Tham, G. Huang, and J. Yan, "Supervised learning with cyclegan for low-dose fdg pet image denoising," *Medical image analysis*, 2020.

[117] Y. Zhou, Z. Yang, H. Zhang, E. I.-C. Chang, Y. Fan, and Y. Xu, "3d segmentation guided style-based generative adversarial networks for pet synthesis," *IEEE TMI*, 2022.

[118] Z. Yang, Y. Zhou, H. Zhang, B. Wei, Y. Fan, and Y. Xu, "Drmc: a generalist model with dynamic routing for multi-center pet image synthesis," in *MICCAI*, 2023.

[119] Y. Luo, L. Zhou, B. Zhan, Y. Fei, J. Zhou, Y. Wang, and D. Shen, "Adaptive rectification based adversarial network with spectrum constraint for high-quality pet image synthesis," *Medical Image Analysis*, 2022.

[120] C. H. McCollough, A. C. Bartley, R. E. Carter, B. Chen, T. A. Drees, P. Edwards, D. R. Holmes III, A. E. Huang, F. Khan, S. Leng *et al.*, "Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge," *Medical physics*, 2017.

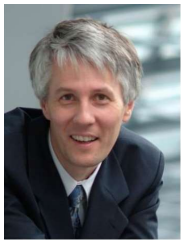
[121] Z. Zhang, L. Yu, X. Liang, W. Zhao, and L. Xing, "Transct: dual-path transformer for low dose computed tomography," in *MICCAI*, 2021.
[122] A. Luthra, H. Sulakhe, T. Mittal, A. Iyer, and S. Yadav, "Eformer: Edge enhancement based transformer for medical image denoising," *arXiv preprint arXiv:2109.08044*, 2021.
[123] D. Wang, F. Fan, Z. Wu, R. Liu, F. Wang, and H. Yu, "Ctformer: convolution-free token2token dilated vision transformer for low-dose ct denoising," *Physics in Medicine & Biology*, 2023.
[124] Ş. Öztürk, O. C. Duran, and T. Çukur, "Denomamba: A fused state-space model for low-dose ct denoising," *arXiv preprint arXiv:2409.13094*, 2024.
[125] "Ixi dataset," 2023, accessed: 2025-08-09. [Online]. Available: http://braindevelopment.org/ixi-dataset/
[126] L. Chen, X. Yang, G. Jeon, M. Anisetti, and K. Liu, "A trusted medical image super-resolution method based on feedback adaptive weighted dense network," *Artificial Intelligence in Medicine*, 2020.
[127] J. Huang, Y. Fang, Y. Wu, H. Wu, Z. Gao, Y. Li, J. Del Ser, J. Xia, and G. Yang, "Swin transformer for fast mri," *Neurocomputing*, 2022.
[128] J. Huang, X. Xing, Z. Gao, and G. Yang, "Swin deformable attention u-net transformer (sdaut) for explainable fast mri," in *MICCAI*, 2022.
[129] H. Sun, Y. Li, Z. Li, R. Yang, Z. Xu, J. Dou, H. Qi, and H. Chen, "Fourier convolution block with global receptive field for mri reconstruction," *Medical Image Analysis*, 2025.
[130] 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.
[131] Y. Wen, T. Gao, J. Zhang, Z. Li, and T. Chen, "Encoder-free multiaxis physics-aware fusion network for remote sensing image dehazing," *TGRS*, 2023.
[132] K. Chi, Y. Yuan, and Q. Wang, "Trinity-net: Gradient-guided swin transformer-based remote sensing image dehazing and beyond," *IEEE TGRS*, 2023.
[133] 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," *TGRS*, 2025.
[134] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE TIP*, 2023.
[135] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater image enhancement via medium transmission-guided multi-color space embedding," *IEEE TIP*, 2021.
[136] Y. Tang, H. Kawasaki, and T. Iwaguchi, "Underwater image enhancement by transformer-based diffusion model with non-uniform sampling for skip strategy," in *ACM MM*, 2023.
[137] C. Zhao, W. Cai, C. Dong, and C. Hu, "Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration," in *CVPR*, 2024.
[138] L. Peng and L. Bian, "Adaptive dual-domain learning for underwater image enhancement," in *AAAI*, 2025.
[139] B. Peng, E. Alcaide, Q. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. Chung, M. Grella *et al.*, "Rwkv: Reinventing rnns for the transformer era," *arXiv preprint arXiv:2305.13048*, 2023.
[140] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *CVPR*, 2017.
[141] S. Zhou, J. Pan, and J. Yang, "Learning an adaptive sparse transformer for efficient image restoration," *IEEE TPAMI*, 2025.

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.

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-Hsuan Yang as a joint training student in 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, Shenzhen Campus, Sun Yat-sen University. His research interests include image processing and 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. His research interests include computational photography and computer vision.



Alois Knoll (Fellow, IEEE) received his diploma (M.Sc.) degree in Electrical/Communications Engineering from the University of Stuttgart, Germany, in 1985 and his Ph.D. (*summa cum laude*) in Computer Science from Technical University of Berlin, Germany, in 1988. He served on the faculty of the Computer Science department at TU Berlin until 1993. He joined the University of Bielefeld, Germany as a full professor and served as the director of the Technical Informatics research group until 2001. Since 2001, he has been a professor at the Department of Informatics, Technical University of Munich (TUM), Germany. He was also on the board of directors of the Central Institute of Medical Technology at TUM (IMETUM). From 2004 to 2006, he was Executive Director of the Institute of Computer Science at TUM. Between 2007 and 2009, he was a member of the EU's highest advisory board on information technology, ISTAG, the Information Society Technology Advisory Group, and a member of its subgroup on Future and Emerging Technologies (FET). His research interests include cognitive, medical and sensor-based robotics, multi-agent systems, data fusion, 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.

Visual-in-Visual: A Unified and Efficient Baseline for Image Restoration—Supplementary Material

Yuning Cui, *Graduate Student Member, IEEE*, Wenqi Ren, *Senior Member, IEEE*, Boxin Shi, *Senior Member, IEEE*, and Alois Knoll, *Fellow, IEEE*

I. Datasets and Implementation Details

The datasets and hyperparameters used in this study are summarized in Table I. We comprehensively evaluate our model on five super-category image restoration tasks: general, all-in-one, composite degradation, ultra-high-definition (UHD), and domain-specific.

Except for the all-in-one setting, where the model is trained on a mixed dataset comprising multiple tasks, we train the model separately for each dataset. Next, we introduce the implementation details for each category.

For general, composite degradation, and UHD image restoration tasks, the model is trained using the Adam optimizer and a dual-domain L_1 loss function [1]–[3]. Following [4], we first pre-train our model on the LFDOF [5] dataset and then fine-tune it on the DPDD [6] dataset for defocus deblurring. For motion deblurring, we adopt a two-stage training paradigm, following the protocol in [7]. Consistent with prior works [8], [9], the quality scores for deraining tasks are computed on the Y channel of the YCbCr color space.

For all-in-one image restoration tasks, our training configurations closely follow those of recent algorithms [10]–[12]. Specifically, our model is trained using the Adam optimizer and a dual-domain L_1 loss function. Noisy images are generated by adding Gaussian noise of levels $\sigma \in \{15, 25, 50\}$ to clean images.

For domain-specific tasks, our training and data processing strategies closely follow those of prior methods within each domain to ensure a fair comparison. No additional training tricks are applied to boost performance.

II. More Visual Comparisons

Figures 1-5 present visual comparisons between the proposed model and competing approaches across different image restoration tasks. In addition, Figures 6 and 7 present additional visual results for real-world image dehazing under varying haze conditions.

Corresponding author: Wenqi Ren (renwq3@mail.sysu.edu.cn)

Yuning Cui 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, China, and also with the State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou, 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.

References

[1] 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 *ICCV*, 2021.
[2] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Revitalizing convolutional network for image restoration," *IEEE TPAMI*, 2024.
[3] ——, "Focal network for image restoration," in *ICCV*, 2023.
[4] L. Ruan, B. Chen, J. Li, and M. Lam, "Learning to deblur using light field generated and real defocus images," in *CVPR*, 2022.
[5] L. Ruan, B. Chen, J. Li, and M.-L. Lam, "Aifnet: All-in-focus image restoration network using a light field-based dataset," *IEEE TCI*, 2021.
[6] A. Abuolaim and M. S. Brown, "Defocus deblurring using dual-pixel data," in *ECCV*, 2020.
[7] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient frequency domain-based transformers for high-quality image deblurring," in *CVPR*, 2023.
[8] X. Chen, H. Li, M. Li, and J. Pan, "Learning a sparse transformer network for effective image deraining," in *CVPR*, 2023.
[9] 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 *CVPR*, 2022.
[10] V. Potlapalli, S. W. Zamir, S. H. Khan, and F. Shahbaz Khan, "Promptir: Prompting for all-in-one image restoration," in *NeurIPS*, 2023.
[11] 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 *ICLR*, 2025.
[12] 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 *CVPR*, 2025.
[13] W.-T. Chen, H.-Y. Fang, J.-J. Ding, C.-C. Tsai, and S.-Y. Kuo, "Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal," in *ECCV*, 2020.
[14] Y.-F. Liu, D.-W. Jaw, S.-C. Huang, and J.-N. Hwang, "Desnownet: Context-aware deep network for snow removal," *IEEE TIP*, 2018.
[15] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *CVPR*, 2017.
[16] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking single-image dehazing and beyond," *IEEE TIP*, 2018.
[17] 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 *ICIP*, 2019.
[18] C. O. Ancuti, C. Ancuti, and R. Timofte, "Nh-haze: An image dehazing benchmark with non-homogeneous hazy and haze-free images," in *CVPRW*, 2020.
[19] Y. Liu, L. Zhu, S. Pei, H. Fu, J. Qin, Q. Zhang, L. Wan, and W. Feng, "From synthetic to real: Image dehazing collaborating with unlabeled real data," in *ACM MM*, 2021.
[20] H. Zhang and V. M. Patel, "Density-aware single image de-raining using a multi-stream dense network," in *CVPR*, 2018.
[21] T. Wang, X. Yang, K. Xu, S. Chen, Q. Zhang, and R. W. Lau, "Spatial attentive single-image deraining with a high quality real rain dataset," in *CVPR*, 2019.
[22] 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.
[23] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE TIP*, 2021.
[24] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, "Deep joint rain detection and removal from a single image," in *CVPR*, 2017.
[25] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE TPAMI*, 2010.

TABLE I
DATASETS AND HYPERPARAMETERS USED FOR FIVE SUPER-CATEGORY IMAGE RESTORATION TASKS. UNLESS OTHERWISE SPECIFIED, THE TRAINING AND TEST SETS ARE FROM THE SAME DATASET.

Setting/Task	Training/Test	Model	Patch	Batch	Learning rate	Iteration
General image restoration						
Image desnowing	SRRS [13]	VIVNet-B/-S	256	8	1e-3	300K
	Snow100K [14]	VIVNet-B/-S	256	8	1e-3	300K
Motion deblurring	GoPro [15]	VIVNet-L	128	64	1e-3	300K
		VIVNet-L	256	16	5e-4	300K
Image dehazing	RESIDE-Indoor [16]/SOTS [16]	VIVNet-S	256	16	1e-3	600K
	Dense-Haze [17]	VIVNet-S	700	4	1e-3	50K
	NH-HAZE [18]	VIVNet-S	700	4	1e-3	50K
	Haze4k [19]	VIVNet-L/-B/-S/-T	256	8	1e-3	300K
Defocus deblurring	LFDOF [5]	VIVNet-L	128	64	1e-3	150K
	DPDD [6]	VIVNet-L	128	64	1e-4	150K
Image deraining	DID-Data [20]	VIVNet-B	256	8	1e-3	300K
	SPA-Data [21]	VIVNet-B	256	8	1e-3	300K
	AGAN-Data [22]	VIVNet-B	256	8	1e-3	300K
Low-light enhancement	LOL-v2 [23]	VIVNet-T	128	32	1e-3	300K
All-in-one image restoration						
Three-task		VIVNet-B	128	32	2e-4	500K
Image deraining	Rain100L [24]					
Image dehazing	RESIDE-β [16]/SOTS [16]					
Image denoising	BSD400 [25], WED [26]/BSD68 [27]					
Five-task (Three-task+2 extra)		VIVNet-B	128	32	2e-4	700K
Motion deblurring	GoPro [15]					
Low-light enhancement	LoLv1 [28]					
Composite degradation						
Two-degradation	LOLBlur [29]	VIVNet-B/-T	256	8/24	1e-3	300K
Three-degradation	CDD-11 [30]	VIVNet-B/-T	256	8/24	1e-3	300K
Ultra-high-definition						
Image dehazing	UHD-Haze [31]	VIVNet-T	768	6	1e-3	100K
Image deraining	4K-Rain13k [32]	VIVNet-T	768	6	1e-3	100K
Image deblurring	UHD-Blur [31]	VIVNet-T	768	6	1e-3	100K
Low-light enhancement	UHD-LL [33]	VIVNet-T	768	6	1e-3	100K
Domain-specific						
Medical						
PET image synthesis	PolarStar M660 [34]	VIVNet-T	128	4	2e-4	300K
CT image denoising	AAPM [35]	VIVNet-T	128	4	2e-4	300K
MRI image super-resolution	IXI [36]	VIVNet-T	128	4	2e-4	300K
Underwater						
Enhancement	LSUI [37]	VIVNet-S	256	16	1e-3	300K
Remote sensing						
Image dehazing	SateHaze1k-thin [38]	VIVNet-S	256	16	1e-3	50K
	SateHaze1k-moderate [38]	VIVNet-S	256	16	1e-3	50K
	SateHaze1k-thick [38]	VIVNet-S	256	16	1e-3	50K

10.62 dB Low-light Image | 21.63 dB Retinexformer | 24.14 dB MambaLLIE | 31.38 dB Ours | PSNR Reference

Fig. 1. Visual comparisons on the LOL-v2-s [23] dataset for low-light image enhancement.

[26] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, H. Li, and L. Zhang, "Waterloo exploration database: New challenges for image quality assessment models," *IEEE TIP*, 2016.

[27] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001.

[28] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *BMVC*, 2018.

[29] S. Zhou, C. Li, and C. Change Loy, "Lednet: Joint low-light enhancement and deblurring in the dark," in *ECCV*, 2022.

[30] Y. Guo, Y. Gao, Y. Lu, H. Zhu, R. W. Liu, and S. He, "Onerestore: A universal restoration framework for composite degradation," in *ECCV*, 2024.

[31] C. Wang, J. Pan, W. Wang, G. Fu, S. Liang, M. Wang, X.-M. Wu, and

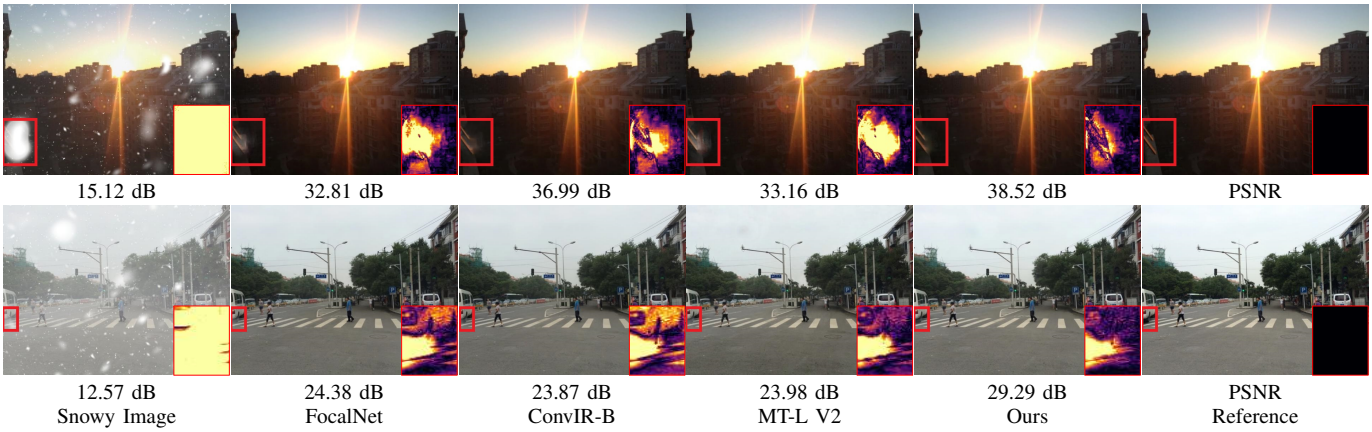


Fig. 2. Visual comparisons on the SRRS [13] dataset for image desnowing. Error heatmaps between restored and reference regions are provided, with brighter pixels indicating larger errors.

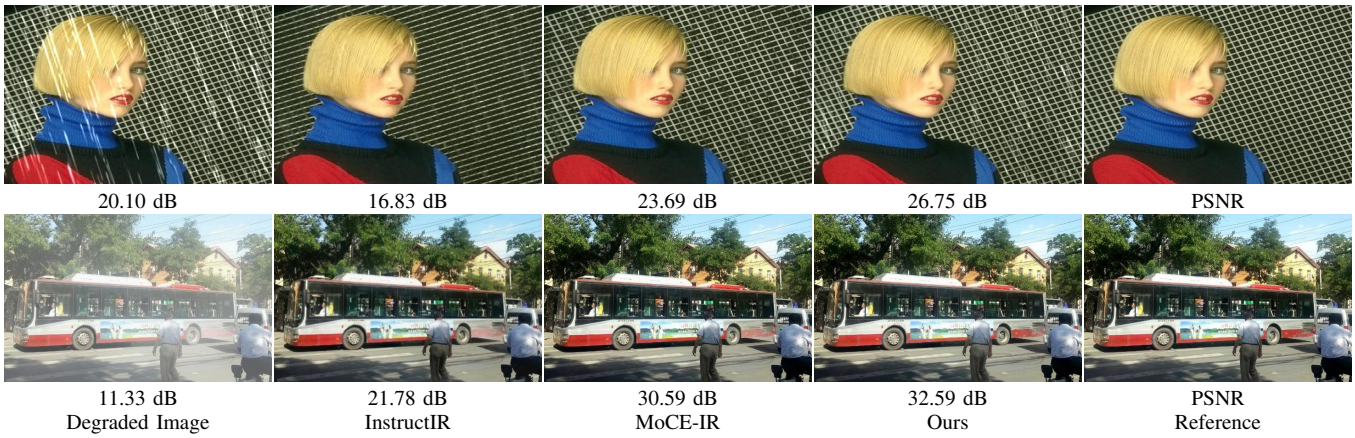


Fig. 3. Visual comparisons under the five-task all-in-one setting. Two examples are obtained from the Rain100L [24] and SOTS-Outdoor [16] datasets for deraining and dehazing, respectively.

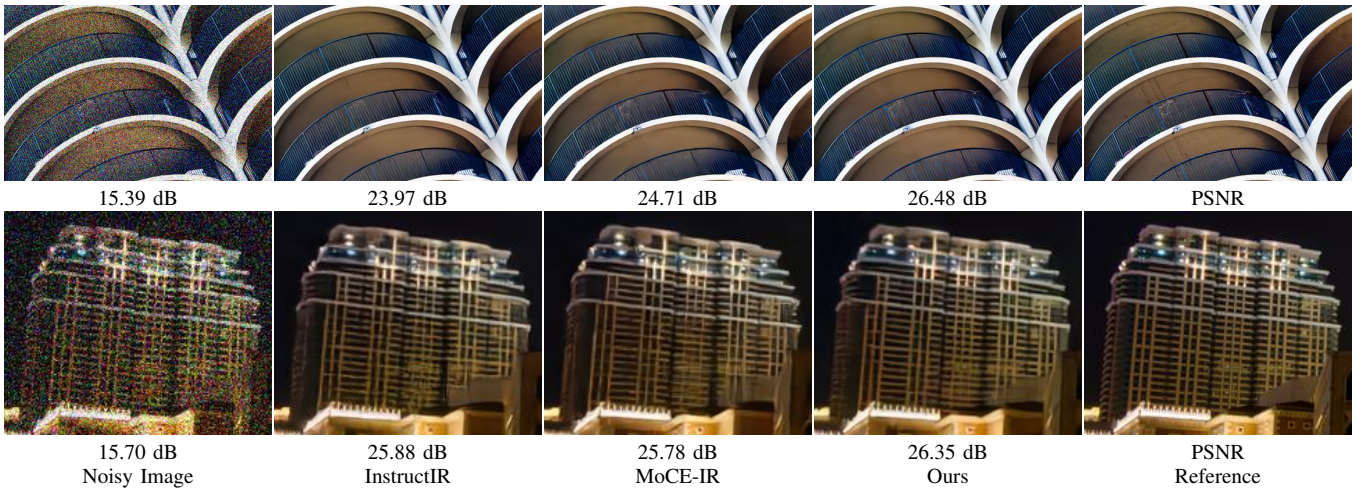


Fig. 4. Visual comparisons for generalization evaluation. The results are obtained by directly applying models pre-trained in the five-task setting to the Urban100 [39] dataset with a noise level of $\sigma = 50$.

J. Liu, "Correlation matching transformation transformers for uhd image restoration," in *AAAI*, 2024.

[32] 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," *arXiv preprint arXiv:2405.17074*, 2024.

[33] C. Li, C.-L. Guo, man zhou, Z. Liang, S. Zhou, R. Feng, and C. C. Loy, "Embedding fourier for ultra-high-definition low-light image enhancement," in *ICLR*, 2023.

[34] Z. Yang, J. Li, H. Zhang, D. Zhao, B. Wei, and Y. Xu, "Restore-rwkv: Efficient and effective medical image restoration with rwkv," *arXiv preprint arXiv:2407.11087*, 2025.

[35] C. H. McCollough, A. C. Bartley, R. E. Carter, B. Chen, T. A. Drees, P. Edwards, D. R. Holmes III, A. E. Huang, F. Khan, S. Leng *et al.*, "Low-dose ct for the detection and classification of metastatic liver lesions: results of the 2016 low dose ct grand challenge," *Medical physics*, 2017.

18.01 dB Haze+Snow | 31.77 dB OneRestore | 32.39 dB MoCE-IR-S | 33.90 dB Ours-B | PSNR Reference

Fig. 5. Visual comparisons on the CDD-11 [30] dataset for image restoration under composite degradations.

Hazy Image | Reference | Ours-S

Fig. 6. Visual results under dense haze conditions [17].

[36] "Ixi dataset," 2023, accessed: 2025-08-09. [Online]. Available: http://braindevelopment.org/ixi-dataset/
[37] L. Peng, C. Zhu, and L. Bian, "U-shape transformer for underwater image enhancement," *IEEE TIP*, 2023.
[38] 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.
[39] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *CVPR*, 2015.

Hazy Image Reference Ours-S

Fig. 7. Visual results on real-world images with varying haze densities [18].