Parameter Efficient Adaptation for Image Restoration with Heterogeneous Mixture-of-Experts

Hang Guo[1] **Tao Dai**[*2] **Yuanchao Bai**[3] **Bin Chen**[3]
Xudong Ren[1] **Zexuan Zhu**[2] **Shu-Tao Xia**[1,4]
[1]Tsinghua University [2]Shenzhen University
[3]Harbin Institute of Technology [4]Peng Cheng Larboratory
https://github.com/csguoh/AdaptIR

Abstract

Designing single-task image restoration models for specific degradation has seen great success in recent years. To achieve generalized image restoration, all-in-one methods have recently been proposed and shown potential for multiple restoration tasks using one single model. Despite the promising results, the existing all-in-one paradigm still suffers from high computational costs as well as limited generalization on unseen degradations. In this work, we introduce an alternative solution to improve the generalization of image restoration models. Drawing inspiration from recent advancements in Parameter Efficient Transfer Learning (PETL), we aim to tune only a small number of parameters to adapt pre-trained restoration models to various tasks. However, current PETL methods fail to generalize across varied restoration tasks due to their homogeneous representation nature. To this end, we propose AdaptIR, a Mixture-of-Experts (MoE) with orthogonal multi-branch design to capture local spatial, global spatial, and channel representation bases, followed by adaptive base combination to obtain heterogeneous representation for different degradations. Extensive experiments demonstrate that our AdaptIR achieves stable performance on single-degradation tasks, and excels in hybrid-degradation tasks, with fine-tuning only 0.6% parameters for 8 hours.

1 Introduction

Image restoration, aiming to restore high-quality images from their degraded counterparts, is a fundamental computer vision problem and has been studied for many years. Due to its ill-posed nature, early research efforts [1, 2, 3] typically focus on developing single-task models, with each model handling only one specific degradation. Consequently, these methods often exhibit limited generalization across different image restoration tasks.

To improve generalization ability, all-in-one image restoration methods [4, 5, 6] have recently been proposed and have attracted great research interest. By training one model with multiple degradation data, these methods enable the single model to handle various degradations. Despite the promising results, the existing all-in-one paradigm still faces several challenges. Firstly, the all-in-one model can only restore degradations encountered during training; once training is complete, the model cannot handle new degradations. Secondly, since the knowledge of restoring multiple degradations is learned by a single model, it incurs a significant cost to train and store these all-in-one models.

In this work, we propose an alternative solution to improve the generalization ability of restoration models in handling multiple degradations. Drawing inspiration from Parameter Efficient Transfer Learning (PETL) [7, 8, 9, 10], we aim to insert a small number of trainable modules into frozen pre-trained restoration backbones. By training only these newly added modules on downstream tasks, the pre-trained restoration backbone can be adapted to unseen restoration tasks. Since only a small

[*]Corresponding author: Tao Dai (daitao.edu@gmail.com).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-0432

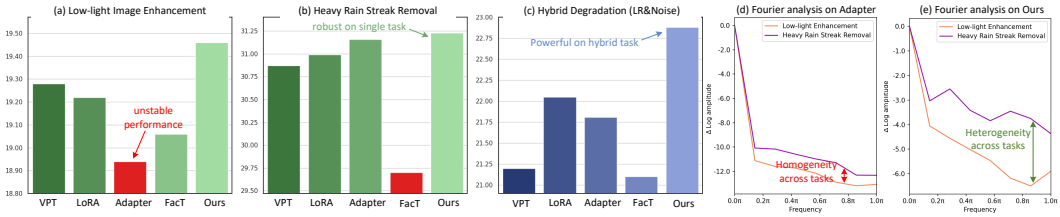


Figure 1: (a)&(b) We find directly applying the current PETL methods to image restoration leads to unstable performance on single degradation. (c) The current PETL method suffers sub-optimal results on hybrid degradation which requires heterogeneous representation. (d)&(e) We use Fourier analysis to visualize Adapter and our AdaptIR and find that Adapter exhibits homogeneous frequency representations even when faced with different degradations, while our AdaptIR can adaptively learn degradation-specific heterogeneous representations. We provide more evidence in Appendix I.

number of parameters need to be trained, the training cost is very small and the training process can converge quickly when new tasks are added. When the training is completed, only the newly added parameters need to be stored, thus greatly reducing the storage cost.

Despite the potential of applying PETL techniques to image restoration, our experiments reveal that existing PETL methods can work normally on specific degradation, but fail to generalize across multiple degradations, exhibiting unstable performance when adapted to different restoration tasks. As shown in Fig. 1(a), the most widely used PETL method, Adapter [8], performs well on the draining task. However, when applying Adapter to the low-light image enhancement task, the adapted model shows significant performance degradation. This phenomenon also occurs with other methods, such as the recent state-of-the-art PETL method FacT [10](Fig. 1(b)). This confusing phenomenon motivates us to discover possible reasons.

To this end, we design preliminary experiments, in which we fine-tune the pre-trained restoration model [11] using existing PETL schemes, and then use Fourier analysis [12] to observe the frequency characteristics of features from these methods. It is observed that the features from current PETL methods exhibit homogeneous representation across different restoration tasks (see Fig. 1(d)). As demonstrated in previous work [6], different restoration tasks prefer certain representations for optimal results, we thus hypothesize that the performance drop occurs when the representation needed to address one specific degradation does not match the homogeneous representation of existing PETL methods. To verify this hypothesis, we further test current PETL methods using the hybrid degradation task (Fig. 1(c)), which requires heterogeneous representations to handle diverse degradations, and we find all existing approaches suffer severe performance drops. Based on the above experiments, we argue that the homogeneous representation of existing PETL methods hinders stable performance on single degradation tasks and advanced performance on hybrid degradation tasks.

In order to learn heterogeneous representations across tasks, one possible solution is the multi-branch structures, where each branch is designed to learn orthogonal representation bases, and then adaptively combine these bases for specific degradation. Following this idea, we propose AdaptIR, a heterogeneous Mixture-of-Experts (MoE) to adapt pre-trained restoration models with heterogeneous representations across tasks. Our AdaptIR adopts orthogonal multi-branch design to learn local spatial, global spatial, and channel representation bases. Specifically, The Local Interaction Module (LIM) employs depth-separable convolution with kernel weight decomposition to exploit local spatial representation. We then employ the Frequency Affine Module (FAM), which performs frequency affine transformation to introduce global spatial modeling ability. Additionally, the Channel Gating Module (CGM) is adopted to capture channel interactions. Finally, we utilize the Adaptive Feature Ensemble to dynamically fuse these three representation bases for specific degradation. Thanks to the heterogeneous representation modeling, our AdaptIR achieves stable performance on single-degradation tasks and advanced performance on hybrid-degradation tasks.

The contributions of our work are as follows: **(i)** We propose a novel PETL paradigm to improve the generalization for image restoration models, and further investigate the specific challenges when applying existing PETL methods to low-level restoration. **(ii)** We introduce AdaptIR, a custom PETL method that employs a heterogeneous MoE for orthogonal representation modeling. To the best of our knowledge, this is the first work to explore parameter-efficient adaptation for image restoration. **(iii)** Experiments on various downstream tasks demonstrate that AdaptIR achieves robust performance on single degradation tasks and advanced results on hybrid degradation tasks.

https://doi.org/10.52202/079017-0432

2 Related Work

2.1 Generalized Image Restoration

Image restoration has attracted a lot of research interest in recent years. Due to the challenging ill-posed nature, some early research paradigms typically study each sub-task in image restoration independently and have recently achieved favorable progress in their respective fields [13, 14, 15, 2, 16]. However, designing such a single-task model is cumbersome, and does not consider the similarities among different tasks. Recently, all-in-one image restoration [4, 5, 6] has offered a way to improve the generalization of image restoration models. By training one single model on multiple degradations, it allows the model to have the ability to handle multiple degradations. For example, AirNet [4] proposes a two-stage training scheme to first learn the degradation representation, which is then used in the following restoration stage. PromptIR [5] utilizes prompt learning to obtain degradation-specific prompts to train the model in an end-to-end manner. Despite the progress, the current all-in-one restoration paradigm can only deal with degradation seen during training and it is inevitable to re-train the model when needing to add new degradations. In addition, incorporating the knowledge of handling multiple degradations into one model has to increase the model size, which causes large training and storage costs.

2.2 Parameter-Efficient Transfer Learning

Parameter efficient transfer learning, which initially came from with NLP [17, 18, 19, 8, 20, 21, 22, 23], aims to catch up with full fine-tuning by training a small number of parameters. Recently, this technique has emerged in the field of computer vision with promising results [9, 7, 24, 10, 25, 26, 27, 28, 29, 30, 31]. For example, VPT [9] adds learnable tokens, also called prompts, to the input sequence of one frozen transformer layer. Adapter [17] employs a bottleneck structure to adapt the pre-trained model. Some attempts also introduce parameterized hypercomplex multiplication layers [22] and re-parameterisation [32] to adapter-based methods. Moreover, LoRA [21] utilizes the low-rank nature of the incremental weight in attention and performs matrix decomposition for parameter efficiency. He et al. [23] go further to identify all the above three approaches from a unified perspective. In addition, NOAH [26] and GLoRA [25] introduce Neural Architecture Search (NAS) to combine different methods. SSF [24] performs a learnable affine transformation on features of the pre-trained model. FacT [10] tensorizes ViT and decomposes the increments into lightweight factors. Although applying PETL methods to pre-trained image restoration models to improve the generalization seems promising, we find that current PETL methods suffer from homogeneous representations when facing different degradations, hindering stable performance across tasks.

3 Method

3.1 Preliminary

In this work, we aim to adapt pre-trained restoration models to multiple downstream tasks by fine-tuning a small number of parameters. Following existing PETL works, we mainly focus on transformer-based restoration models since transformer has been shown to be suitable for pre-training [33] and there is no CNN-based pre-trained model available. As shown in Fig. 2, a typical pre-trained restoration model [11, 34] usually contains one large transformer body as well as task-specific heads and tails. Given the pre-assigned task type, the low-quality image I_{LQ} will first go through the corresponding head to get the shallow feature X_{head}. After that, X_{head} is flattened into a 1D sequence on the spatial dimension and is input to the transformer body which contains several stacked transformer blocks with each block containing multiple transformer layers [35]. Finally, a skip connection is adopted followed by the task-specific tail to reconstruct the high-quality image I_{HQ}. During the pre-training stage, gradients from multiple tasks are used to update the shared body as well as the corresponding task-specific head and tail. After pre-training the restoration model, previous common practice fine-tunes all parameters of the pre-trained model for specific downstream tasks, which burdens training and storage due to the per-task model weights.

https://doi.org/10.52202/079017-0432

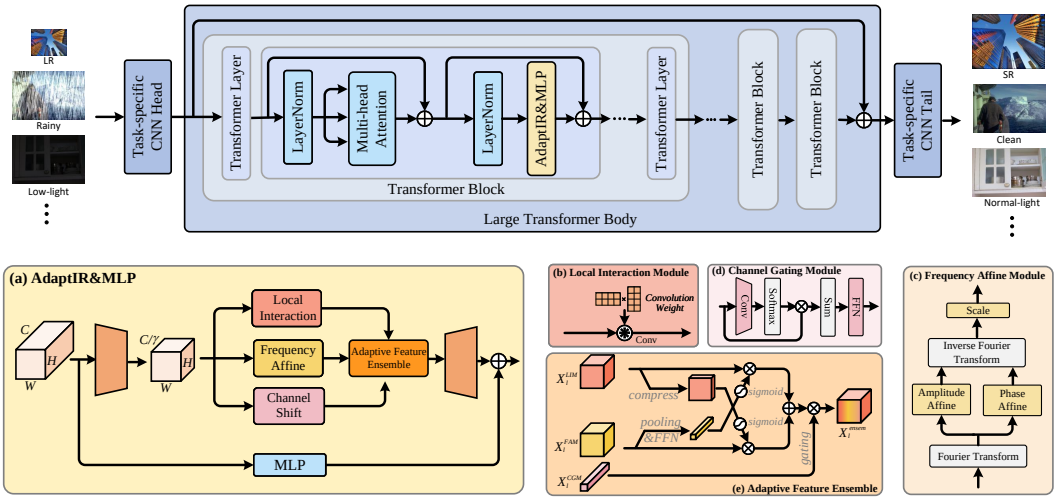


Figure 2: An illustration of the proposed AdaptIR. Our AdaptIR is placed parallel to the frozen MLP in one transformer layer and thus can be seamlessly inserted into various transformer-based pre-trained restoration models.

3.2 Heterogeneous Representation Learning

To obtain stable performance across multiple restoration tasks, it is crucial to allow the learning of heterogeneous representation for different degradations. To this end, we formalize AdaptIR as a multi-branch MoE structure, where each branch learns representations orthogonal to each other to form the representation bases, and then these bases are adaptively combined to achieve degradation-specific representation. Formally, as shown in Fig. 2(a), since the transformer body flattens the l-th layer feature into a 1D token sequence, we first restore the 2D image structure to obtain $X_l \in \mathbb{R}^{C\times H\times W}$. After that, we apply the 1×1 convolution with channel reduction rate γ to transfer X_l to low-dimension space for parameter efficiency and obtain the intrinsic feature $X_l^{intrin} \in \mathbb{R}^{\frac{C}{\gamma}\times H\times W}$. Then three parallel branches are orthogonally designed to learn local spatial, global spatial, and channel bases. Next, bases from these three branches are adaptively ensembled to obtain representation X_l^{adapt} for specific degradation. Finally, X_l^{adapt} is added to the output of the frozen MLP to adapt the pre-trained restoration models. Details of the three branches are given below.

Local Interaction Module. We first introduce the Local Interaction Module (LIM) to model the local spatial representation. As shown in Fig. 2(b), the proposed LIM is implemented by the depth-wise convolution with weight factorization for parameter efficiency. Specifically, given the convolution weight $W \in \mathbb{R}^{C_{in}\times \frac{C_{out}}{group}\times K\times K}$, where C_{in}, C_{out} are input and output channel, K is the kernel size and $group$ is the number of convolution groups, we first reshape W into a 2D weight matrix $W' \in \mathbb{R}^{C_{in}\times \frac{C_{out}}{group}K^2}$, and then decompose W' into multiplication of two low-rank weight matrices:

$$W' = UV^\top, \tag{1}$$

where $U \in \mathbb{R}^{C_{in}\times r}$, $V \in \mathbb{R}^{\frac{C_{out}}{group}K^2\times r}$ and r is the rank to trade-off performance and efficiency. Then we reshape W' to the original kernel size and use it to convolve X_l^{intrin} to get X_l^{LIM}:

$$X_l^{LIM} = \mathrm{Reshape}(W') \circledast X_l^{intrin}, \tag{2}$$

where $\circledast$ denotes convolution operator, and $\mathrm{Reshape}(\cdot)$ transforms 2D matrices into 4D convolution kernel weights.

Frequency Affine Module. We then consider modeling global spatial to achieve orthogonal spatial modeling to LIM. A possible solution is to introduce the attention mechanism [35] which has a global receptive field. However, the attention comes at the cost of high complexity, which goes against the principle of parameter efficiency. In this work, we resort to the frequency domain for a solution. Specifically, we apply the Fast Fourier Transform (FFT) on X_l^{intrin} to obtain the corresponding

frequency feature map $X_l^{\mathcal{F}} \in \mathbb{C}^{\frac{C}{\gamma} \times H \times (\lfloor \frac{W}{2} \rfloor + 1)}$:

$$X_l^{\mathcal{F}}(u,v) = \frac{1}{HW} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X_l^{intrin}(h,w) e^{-2\pi i (\frac{uh}{H} + \frac{vw}{W})}, \tag{3}$$

As can be seen from Eq.3, a good property of FFT is that each position of the frequency feature map is the weighted sum of all features in the spatial domain. Therefore, performing pixel-wise projection on $X_l^{\mathcal{F}}$ is equivalent to performing a global operator in the spatial domain.

Motivated by this observation, we propose the Frequency Affine Module (FAM) to take advantage of the inherent global representation in $X_l^{\mathcal{F}}$ (see Fig. 2(c)). Concretely, we perform the affine transformation on amplitude map Mag_l and phase map Pha_l respectively with depth-separable 1×1 convolution. To ensure numerical stability during the early training stages, we initialize the transformation layers as all-one weights and zero bias. Subsequently, the inverse Fast Fourier Transform (iFFT) is applied to convert the affined feature back to the spatial domain. Finally, another depth-separable 1×1 convolution is used as a scale layer for subsequent feature ensemble. In short, the whole process can be formalized as:

$$\begin{aligned} &[Mag_l, Pha_l] = \mathrm{FFT}(X_l^{intrin}), \\ &X_l^{FAM} = \mathrm{Conv}(\mathrm{iFFT}(\mathrm{to\_complex}(\phi_1(Mag_l), \phi_2(Pha_l)))), \end{aligned} \tag{4}$$

where $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are the frequency projection function and $\mathrm{to\_complex}(\cdot,\cdot)$ converts the magnitude and phase to complex numbers.

Channel Gating Module. The above LIM and FAM both adopt the depth-separable strategy for parameter efficiency. To allow for another orthogonal representation, we further develop the Channel Gating Module (CGM) for salient channel selection. As shown in Fig. 2(d), we first obtain the spatial weight mask $\mathcal{M}_l \in \mathbb{R}^{1 \times H \times W}$ by employing 1×1 convolution which compresses the channel dimension of X_l^{intrin} to 1, followed by the $\mathrm{Softmax}$ on the spatial dimension:

$$\mathcal{M}_l = \mathrm{Softmax}(\mathrm{Conv}(X_l^{intrin})). \tag{5}$$

We then apply $\mathcal{M}_l$ on each channel of X_l^{intrin} to perform spatially weighted summation to obtain the channel vector which will go through a Feed Forward Network (FFN) to generate the channel gating factor $X_l^{CGM} \in \mathbb{R}^{\frac{C}{\gamma} \times 1 \times 1}$:

$$X_l^{CGM} = \mathrm{FFN}(\sum_{h,w} \mathcal{M}_l \otimes X_l^{intrin}), \tag{6}$$

where $\otimes$ denotes the Hadamard product.

Adaptive Feature Ensemble. The orthogonal modeling from local spatial, global spatial, and channel can serve as favorable representation bases, and we further introduce the Adaptive Feature Ensemble to learn their combinations to obtain degradation-specific representations. As shown in Fig. 2(e), we use convolution to compress the channel of X_l^{LIM} to 1 to extract its spatial details, while applying global average pooling and FFN on X_l^{FAM} to preserve the global information. Next, the sigmoid activation is employed to generate dynamic weights for multiplication and produces the adaptive spatial features $X_l^{spatial}$. After that, we use X_l^{CGM} to perform channel selection on $X_l^{spatial}$ to obtain the degradation-specific ensemble feature $X_l^{ensem} \in \mathbb{R}^{\frac{C}{\gamma} \times H \times W}$.

At last, a 1×1 convolution is employed to up-dimension the X_l^{ensem} to generate $X_l^{adapt} \in \mathbb{R}^{C \times H \times W}$. For stability, we use zero to initialize the convolution weights. Then, the X_l^{adapt} is added to the output of frozen MLP as residual to adapt pre-trained models to downstream tasks.

3.3 Parameter Efficient Training

During training, we freeze all parameters in the pre-trained model, including task-specific heads and tails as well as the transformer body except for the proposed AdaptIR. A simple L_1 loss is employed to provide pixel-level supervision:

$$\mathcal{L}_{pix} = ||I_{HQ} - I_{LQ}||_1, \tag{7}$$

where $||\cdot||_1$ denotes L_1 norm.

https://doi.org/10.52202/079017-0432

4 Experiment

We first employ single-degradation restoration tasks to assess the performance stability of different PETL methods, including image SR, color image denoising, image deraining, and low-light enhancement. Subsequently, we introduce hybrid degradation to further evaluate the ability to learn heterogeneous representations. In addition, we compare with recent all-in-one methods in both effectiveness and efficiency to demonstrate the advantages of applying PETL for generalized image restoration. Finally, we conduct ablation studies to reveal the working mechanism of the proposed method as well as different design choices. Since evaluating the performance stability requires experiments on multiple single-degradation tasks, due to the page limit, the related experiments can be seen in the Appendix E.1.

4.1 Experimental Settings

Datasets. For image SR, we choose DIV2K [36] and Flickr2K [37] as the training set, and we evaluate on Set5 [38], Set14 [39], BSDS100 [40], Urban100 [41], and Manga109 [42]. For color image denoising, training sets consist of DIV2K [36], Flickr2K [37], BSD400 [40], and WED [43], and we have two testing sets: CBSD68 [44] and Urban100 [41]. For image deraining, we evaluate using the Rain100L [15] and Rain100H [15] benchmarks, corresponding to light/heavy rain streaks. For evaluation on hybrid degradation, where one image contains multiple degradation types, we choose two representatives, consisting of low-resolution and noise as well as low-resolution and JPEG artifact compression, and we add noise or apply JEPG compression on the low-resolution images to synthesize second-order degraded images. For low-light image enhancement, we utilize the training and testing set of LOLv1 [45].

Evaluation Metrics. We use the PSNR and SSIM to evaluate the effectiveness. The PSNR/SSIM of image SR, deraining, and second-order degradation are computed on the Y channel from the YCbCr space, and we evaluate the RGB channel for denoising and low-light image enhancement. Moreover, we use trainable #param to measure efficiency.

Baseline Setup. This work focuses on transferring pre-trained restoration models to downstream tasks under low parameter budgets. Since there is little work studying PETL on image restoration, we reproduce existing PETL approaches and compare them with the proposed AdaptIR. Specifically, we include the following representative PETL methods: **i)** VPT [9], where the learnable prompts are inserted as the input token of transformer layers, and we compare VPT_{Deep} [9] in experiments because of its better performance. **ii)** Adapter [17], which introduces bottleneck structure placed after Attention and MLP. **iii)** LoRA [21], which adds parallel sub-networks to learn low-rank incremental matrices of query and value. **iv)** AdaptFormer [7], which inserts a tunable module parallel to MLP. **v)** SSF [24], where learnable scale and shift factors are used to modulate the frozen features. **vi)** FacT [10], which tensorises a ViT and then decomposes the incremental weights. We also present results of **vii)** full fine-tuning (Full-ft), and **viii)** directly applying pre-trained models to downstream tasks (Pretrain), to provide more insights. For readers unfamiliar with PETL, we have also provided a basic background introduction in Appendix A.

Implementation Details. We use two pre-trained transformer-based restoration models, *i.e.*, IPT [11] and EDT [34], as the base models to evaluate different PETL methods. We control tunable parameters by adjusting channel reduction rate γ. We use AdamW [46] as the optimizer and train for 500 epochs. The learning rate is initialized to 1e-4 and decayed by half at {250,400,450,475} epochs. All experiments are conducted on four NVIDIA 3080Ti GPUs.

4.2 Comparison on Hybrid Degradation Tasks

In order to obtain convincing evaluation results, it is tedious and time-consuming to observe the stability of one particular PETL method on multiple single-degradation tasks. Here, we introduce hybrid-degradation restoration. Since restoring hybrid degraded images requires a heterogeneous representation of the PETL methods, and thus the hybrid degradation is more suitable for evaluation.

In this work, we consider the second-order degradation as a representative of hybrid degradation. Specifically, we employ two different types of second-order degradations, *i.e.*, the $\times 4$ low-resolution and noise with σ=30 (denoted as LR4&Noise30) as well as the $\times 4$ low-resolution and JEPG compression with quality factor q=30 (denoted as LR4&JPEG30). Moreover, we also include the classic

Table 1: Quantitative comparison for hybrid-degradation restoration tasks. The best and the second best results are in red and blue.

Method	Degradation	#param	Set5		Set14		BSDS100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Full-ft	LR4&Noise30	119M	27.24	0.7859	25.56	0.6686	25.02	0.6166	24.02	0.6967	26.31	0.8245
Pretrain	LR4&Noise30	-	19.74	0.3569	19.27	0.3114	19.09	0.2783	18.54	0.3254	19.75	0.3832
SSF [24]	LR4&Noise30	373K	25.41	0.6720	24.02	0.5761	24.06	0.5411	21.89	0.5514	23.33	0.6736
VPT [9]	LR4&Noise30	884K	24.11	0.5570	22.97	0.4722	22.91	0.4336	21.20	0.4527	22.61	0.5570
Adapter [8]	LR4&Noise30	691K	25.60	0.6862	24.16	0.5856	24.17	0.5498	22.05	0.5640	23.61	0.6904
LoRA [21]	LR4&Noise30	995K	25.19	0.6371	23.82	0.5405	23.82	0.5026	21.81	0.5193	23.30	0.6396
Adaptfor. [7]	LR4&Noise30	677K	26.10	0.7138	24.58	0.6095	24.44	0.5686	22.52	0.5976	24.38	0.7296
FacT [10]	LR4&Noise30	537K	25.70	0.6963	24.24	0.5944	24.25	0.5586	21.10	0.5727	23.63	0.6993
MoE	LR4&Noise30	667K	26.35	0.7335	24.80	0.6254	24.59	0.5835	22.77	0.6188	24.73	0.7517
Ours	LR4&Noise30	697K	26.48	0.7441	24.88	0.6345	24.67	0.6279	22.88	0.5932	24.96	0.7625
Full-ft	LR4&JPEG30	119M	27.21	0.7778	25.49	0.6563	25.08	0.6076	23.54	0.6687	25.48	0.7971
Pretrain	LR4&JPEG30	-	25.23	0.6702	24.12	0.5917	24.19	0.5627	21.74	0.5654	22.93	0.6732
SSF [24]	LR4&JPEG30	373K	26.26	0.7321	24.81	0.6285	24.71	0.5882	22.44	0.6085	23.92	0.7350
VPT [9]	LR4&JPEG30	884K	26.63	0.7497	25.14	0.6414	24.89	0.5974	22.96	0.6377	24.53	0.7591
Adapter [8]	LR4&JPEG30	691K	26.73	0.7554	25.22	0.6448	24.92	0.5999	23.09	0.6447	24.74	0.7677
LoRA [21]	LR4&JPEG30	995K	26.64	0.7501	25.17	0.6424	24.91	0.5983	23.02	0.6405	24.64	0.7619
Adaptfor. [7]	LR4&JPEG30	677K	26.74	0.7562	23.08	0.6441	25.22	0.6447	24.92	0.5996	24.72	0.7669
FacT [10]	LR4&JPEG30	537K	26.71	0.7557	25.22	0.6450	24.93	0.5998	23.08	0.6446	24.74	0.7681
MoE	LR4&JPEG30	667K	26.80	0.7590	25.26	0.6465	24.04	0.6009	23.14	0.6477	24.81	0.7708
Ours	LR4&JPEG30	697K	26.91	0.7646	25.34	0.6502	24.98	0.6032	23.25	0.6541	25.02	0.7791

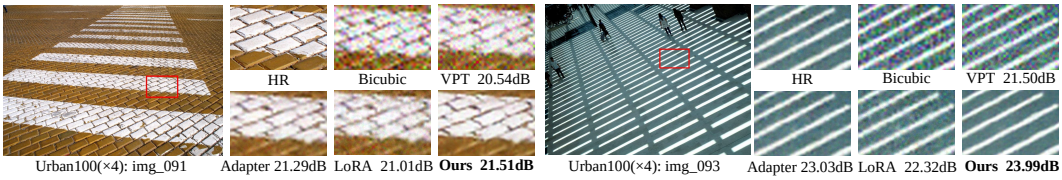


Figure 3: Visual comparison on hybrid degradation with LR4&Noise30. We provide more visualization in Appendix E.1.

MoE [47, 48, 49], which also employs the multi-branch structure but the design of each branch is the same, to give the impact of the multi-branch structure on the performance.

Tab. 1 gives the results. Consistent with the previous analysis in Fig. 1, existing PETL methods suffer severe performance drops on hybrid degradation tasks due to the difficulty of learning heterogeneous representations. Interestingly, even the simple MoE baseline which only uses the multi-branch structure outperforms the current state-of-the-art PETL methods, suggesting that multi-branch structures are promising for heterogeneity across tasks. However, since each branch of the classical MoE employs the same structure, it struggles to capture orthogonal representation bases from different branches. In contrast, our method achieves consistent state-of-the-art performance across all tasks and on all datasets. For example, our AdaptIR outperforms the state-of-the-art PETL method FacT [10] by 1.78dB on Urban100 with LR4&Noise30, and 0.28dB on Manga109 with LR4&JEPG30. By orthogonally designing branches to obtain representation bases and then adaptively combining them, our AdaptIR allows for heterogeneous representations across different tasks. We also give several visual results in Fig. 3, and our AdaptIR can well handle complex degradation.

4.3 Comparison with All-in-One Methods

Recently, all-in-one image restoration methods [5, 4], which learn a single restoration model for various degradations, have shown to be a promising paradigm in achieving generalized image restoration. Here, we compare our AdaptIR with these methods on both single-task and multi-task setups in Tab. 2. For the single-task setting, our method achieves better PSNR results, *e.g.* 0.31dB higher than PromptIR on denoising σ=50. In addition, the performance advantage of our AdaptIR still preserves the multi-task setup. For instance, our AdaptIR outperforms PromptIR by even 4.9dB PSNR and 0.016 SSIM on light rain streak removal. This is because all-in-one methods need to learn multiple degradation restoration within one model, resulting in learning difficulties, and the problem of negative transfer among different tasks [50] can also lead to performance degradation. By contrast, the heterogeneous representation from the orthogonal design facilitates the stable performance of our

https://doi.org/10.52202/079017-0432

Table 2: Comparison with all-in-one image restoration methods under single-task setting. The 'training time' of AdaptIR refers to the downstream fine-tuning time excluding the pre-training stage.

Method	task	dataset	#param	training time	GPU memory	PSNR	SSIM
AirNet [4]	light derain	Rain100L	8.75M	∼48h	∼11G	34.90	0.977
PromptIR [5]	light derain	Rain100L	97M	∼84h	∼128G	37.04	0.979
Ours	light derain	Rain100L	697K	∼8h	∼8G	37.81	0.981
AirNet [4]	denoise σ=50	Urban100	8.75M	∼48h	∼11G	28.88	0.871
PromptIR [5]	denoise σ=50	Urban100	97M	∼84h	∼128G	29.39	0.881
Ours	denoise σ=50	Urban100	697K	∼8h	∼8G	29.70	0.881

Table 3: Comparison with all-in-one image restoration methods under multi-task setting.

Method	#param	GPU memory	training time	light derain	denoise σ=25	denoise σ=30
AirNet [4]	8.7M	∼11G	∼48h	34.90/0.967	31.90/0.914	28.68/0.861
PromptIR [5]	97.1M	∼128G	∼48h	36.37/0.972	32.09/0.919	28.99/0.871
Ours	697K	∼8G	∼10h	41.27/0.988	32.64/0.926	29.16/0.875

Figure 4: Fouriur analysis on outputs from LIM and FAM.

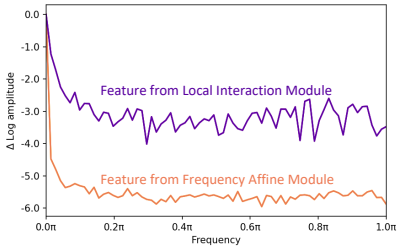
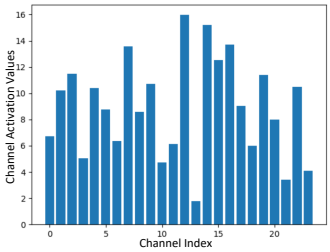


Figure 5: Channel activation visualization on outputs from CGM.



AdaptIR across different degradations. As for efficiency, our AdaptIR only trains 0.7% parameters than that of PromptIR with a fast fine-tuning process. We provide a detailed summarization and discussion about the existing multi-task restoration paradigm in Appendix C.

4.4 Discussion

Why does the Proposed Methods Work? Our proposed AdaptIR adopts the heterogeneous MoE structure to allow diverse representation learning. Here, we delve deep to verify whether the model design can indeed influence the learned features. For LIM and FAM, we visualize the frequency characteristics of their outputs in Fig. 4. It can be seen that LIM's relative log amplitude at π is 11.02 higher than FAM, suggesting it has learned to capture high-frequency local textures. Meanwhile, more than 95% of energy is centralized within 0.05π for FAM, indicating it can well model low-frequency global structure. For CGM, we visualize the channel activation in Fig. 5, and find large activation differences across channels, with a large variance of 96.10, indicating that the CGM learns to select degradation-specific channels.

Scaling Trainable Parameters. We compare the performance of different PETL methods under varying parameter budgets. We use the hybrid degradation LR4&Noise30 in this setup. Fig. 6 shows the results. It can be seen that the proposed method surpasses other strong baselines across various parameter settings, demonstrating the strong scalability of the proposed method.

How About the Performance on Other Pre-trained Models? The above experiments employ IPT [11] as the base model. In order to verify the generalization of the proposed method, we further adopt another pre-trained image restoration model EDT [34] as the frozen base model. Tab. 4 represents the results. It can be seen that the proposed method maintains state-of-the-art performance by tuning only 1.5% parameters. More experiments with EDT can be seen in Appendix E.2.

4.5 Ablation Study

Parameter Efficient Designs. In this work, we introduce several techniques to achieve orthogonal representation learning. Here, we ablate to study the impact of these choices. The results, presented in Tab. 5, indicate that (1) the adaptive feature ensemble can assemble representations according to specific degradation, without which will cause a performance drop. In addition, (2)&(3) removing the

https://doi.org/10.52202/079017-0432

Table 4: Comparison on generalization ability with more pretrained base model.

Method	#param	Set5	Set14	BSDS 100	Urban 100	Manga 109
Full-ft	11.6M	27.32	25.60	25.03	24.10	26.42
Pretrain	-	19.29	18.45	18.27	17.92	19.25
SSF [24]	117K	26.92	25.24	24.83	23.41	25.77
VPT [9]	311K	24.19	22.91	22.81	21.12	22.49
Adapter [17]	194K	26.92	25.27	24.81	23.48	25.84
LoRA [21]	259K	26.91	25.25	24.80	23.46	25.81
AdaptFor. [7]	162K	26.99	25.31	24.85	23.59	25.95
FacT [10]	174K	26.89	25.25	24.81	23.43	25.78
Ours	173K	27.04	25.34	24.87	23.60	25.97

Figure 6: Scalability comparison with different PETL methods.

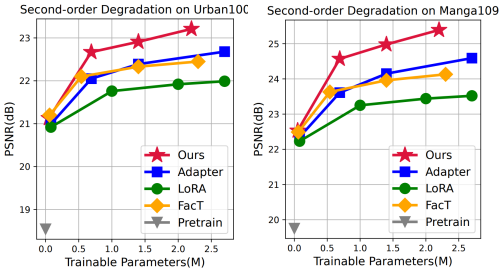


Table 5: Ablation of different design choices on PSNR(dB). 'Baseline' refers to the setting of depth-separable projection in LIM and FAM, as well as the channel-spatial orthogonal modeling.

Settings	#param	Set5	Set14	Urban100
(0)Baseline	697K	26.48	24.88	22.88
(1)+w/o adaptive feature ensemble	692K	24.26	24.88	22.82
(2)+w/o depth-separable in LIM	718K	25.67	24.22	22.10
(3)+w/o depth-separable in FAM	728K	25.73	24.28	22.17
(4)+w/o CGM&w/o depth-separable	743K	24.26	23.15	21.36

Table 6: Ablation experiments of different components on PSNR(dB).

LIM	FAM	CGM	#param	Set5	Set14	Urban100
		✔	680K	23.64	22.66	20.97
	✔	✔	682K	25.52	24.17	22.05
✔	✔		678K	26.11	24.60	22.52
✔	✔	✔	697K	26.48	24.88	22.88

Table 7: Ablation for different insertion positions and forms on PSNR(dB).

position	form	Set5	Set14	Urban100
MLP	parallel	26.48	24.88	22.88
Attention	parallel	26.28	24.70	22.59
MLP	sequential	26.35	24.77	22.67
Attention	sequential	25.60	24.19	22.07

depth-separable design in LIM or FAM will conflict with the channel modeling in CGM, and lead to sub-optimal results. Further, (4) removing the CGM branch while allowing full-channel interaction in other branches results in poor performance, which we attribute to the learning difficulty of modeling channel and spatial simultaneously.

Ablation for Components. In the proposed AdaptIR, three parallel branches are developed to learn orthogonal bases. We ablate to discern the roles of different branches. As shown in Tab. 6, separate utilization of one or two branches only yields sub-optimal results owing to the insufficient representation. And the combination of the three branches achieves the best results.

Insertion Position and Form. There are various options for both the insertion location and form of our AdapIR. The impact of these choices is shown in Tab. 7. It can be seen that inserting AdaptIR into MLP achieves better performance under both parallel and sequential forms. This is because there is a certain dependency between the well-trained MLP and attention, and insertion into the middle of them will damage this relationship. Moreover, the parallel insertion form performs better than its sequential counterpart. We argue that parallel form can preserve the knowledge of frozen features through summation, thus reducing the learning difficulty.

5 Conclusion

In this work, we explore for the first time the potential of parameter-efficient adaptation to improve the generalization of image restoration models. We observe that current PETL methods struggle to generalize to multiple single-degradation tasks and suffer from performance degradation on hybrid-degradation tasks. We identify that this issue arises from the misalignment between the degradation-required representation and the homogeneity in current PETL methods. Based on this observation, we propose AdaptIR, a heterogeneous Mixture-of-Experts (MoE) to learn local spatial, global spatial, and channel orthogonal bases under low parameter budgets, followed by the adaptive feature ensemble to dynamically fuse these bases for degradation-specific representation. Extensive experiments validate our AdaptIR as a versatile and powerful adaptation solution.

https://doi.org/10.52202/079017-0432

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309,62171248), Shenzhen Science and Technology Program (JCYJ20220818101014030, JCYJ20220818101012025), and the PCNL KEY project (PCL2023AS6-1).

Bibliography

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *ECCV*, pages 184–199. Springer, 2014. 1

[2] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE TIP*, 26(7):3142–3155, 2017. 1, 3

[3] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, pages 3937–3946, 2019. 1

[4] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17452–17462, 2022. 1, 3, 7, 8, 14

[5] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one blind image restoration. *arXiv preprint arXiv:2306.13090*, 2023. 1, 3, 7, 8, 13, 14

[6] Dongwon Park, Byung Hyun Lee, and Se Young Chun. All-in-one image restoration for unknown degradations using adaptive discriminative filters for specific degradations. In *CVPR*, pages 5815–5824. IEEE, 2023. 1, 2, 3

[7] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. AdaptFormer: Adapting vision transformers for scalable visual recognition. *NeurIPS*, 35:16664–16678, 2022. 1, 3, 6, 7, 9, 13, 15, 16

[8] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1, 2, 3, 7, 18

[9] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 1, 3, 6, 7, 9, 13, 15, 16

[10] Shibo Jie and Zhi-Hong Deng. FacT: Factor-tuning for lightweight adaptation on vision transformer. In *AAAI*, volume 37, pages 1060–1068, 2023. 1, 2, 3, 6, 7, 9, 13, 15, 16, 17, 18

[11] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *CVPR*, pages 12299–12310, 2021. 2, 3, 6, 8, 15, 16

[12] Namuk Park and Songkuk Kim. How do vision transformers work? In *ICLR*, 2021. 2

[13] Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep CNN denoiser prior for image restoration. In *CVPR*, pages 3929–3938, 2017. 3

[14] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *CVPR*, pages 11065–11074, 2019. 3, 15

[15] Wenhan Yang, Robby T Tan, Jiashi Feng, Zongming Guo, Shuicheng Yan, and Jiaying Liu. Joint rain detection and removal from a single image with contextualized deep networks. *IEEE TPAMI*, 42(6):1377–1393, 2019. 3, 6, 19

[16] Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. *arXiv preprint arXiv:2402.15648*, 2024. 3

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. pages 2790–2799. PMLR, 2019. 3, 6, 9, 13, 15, 16

[18] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021. 3

[19] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 3

[20] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 3, 6, 7, 9, 13, 15, 16, 17, 18

[22] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. *NeurIPS*, 34:1022–1035, 2021. 3

[23] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021. 3

[24] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning. *NeurIPS*, 35:109–123, 2022. 3, 6, 7, 9, 13, 15, 16

[25] Arnav Chavan, Zhuang Liu, Deepak Gupta, Eric Xing, and Zhiqiang Shen. One-for-All: Generalized LoRA for parameter-efficient fine-tuning. *arXiv preprint arXiv:2306.07967*, 2023. 3

[26] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. 3

[27] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 3

[28] Yen-Cheng Liu, Chih-Yao Ma, Junjiao Tian, Zijian He, and Zsolt Kira. Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks. *NeurIPS*, 35:36889–36901, 2022. 3

[29] Shibo Jie and Zhi-Hong Deng. Convolutional bypasses are better vision transformer adapters. *arXiv preprint arXiv:2207.07039*, 2022. 3

[30] Shibo Jie, Haoqing Wang, and Zhi-Hong Deng. Revisiting the parameter efficiency of adapters from the perspective of precision redundancy. In *ICCV*, pages 17217–17226, 2023. 3

[31] Taolin Zhang, Jinpeng Wang, Hang Guo, Tao Dai, Bin Chen, and Shu-Tao Xia. BoostAdapter: Improving test-time adaptation via regional bootstrapping. *arXiv preprint arXiv:2410.15430*, 2024. 3

[32] Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*, 2023. 3

[33] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[34] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 3, 6, 8, 15, 16

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3, 4, 13

[36] Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *CVPRW*, pages 126–135, 2017. 6

[37] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. NTIRE 2017 challenge on single image super-resolution: Methods and results. In *CVPRW*, pages 114–125, 2017. 6

[38] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012. 6

[39] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Curves and Surfaces: 7th International Conference, Avignon, France, June 24-30, 2010, Revised Selected Papers 7*, pages 711–730. Springer, 2012. 6

[40] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE TPAMI*, 33(5):898–916, 2010. 6

[41] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, pages 5197–5206, 2015. 6

 https://doi.org/10.52202/079017-0432

[42] Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based manga retrieval using Manga109 dataset. *Multimedia Tools and Applications*, 76:21811–21838, 2017. 6

[43] Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo exploration database: New challenges for image quality assessment models. *IEEE TIP*, 26(2):1004–1016, 2016. 6

[44] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *ICCV*, volume 2, pages 416–423. IEEE, 2001. 6

[45] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 6, 19

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6, 16

[47] Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*, 2021. 7

[48] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *NeurIPS*, 35:2664–2678, 2022. 7

[49] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *NeurIPS*, 34:8583–8595, 2021. 7

[50] Wenlong Zhang, Xiaohui Li, Guangyuan Shi, Xiangyu Chen, Yu Qiao, Xiaoyun Zhang, Xiao-Ming Wu, and Chao Dong. Real-world image super-resolution as multi-task learning. *NeurIPS*, 36, 2024. 7

[51] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 13

[52] Yihao Liu, Xiangyu Chen, Xianzheng Ma, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Unifying image processing as visual prompting question answering. *arXiv preprint arXiv:2310.10513*, 2023. 13, 14

[53] Jiaqi Ma, Tianheng Cheng, Guoli Wang, Qian Zhang, Xinggang Wang, and Lefei Zhang. ProRes: Exploring degradation-aware visual prompt for universal image restoration. *arXiv preprint arXiv:2306.13653*, 2023. 13, 14

[54] Cong Wang, Jinshan Pan, Wei Wang, Jiangxin Dong, Mengzhu Wang, Yakun Ju, and Junyang Chen. Promptrestorer: A prompting image restoration method with degradation perception. *NeurIPS*, 36:8898–8912, 2023. 13, 14

[55] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, pages 286–301, 2018. 15

[56] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 15

[57] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE TIP*, 27(9):4608–4622, 2018. 15

[58] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *CVPR*, pages 2472–2481, 2018. 15

[59] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 16

[60] Wei Chen, Wang Wenjing, Yang Wenhan, and Liu Jiaying. Deep retinex decomposition for low-light enhancement. In *BMVC*, 2018. 16

[61] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *CVPR*, pages 2281–2290, 2020. 16

Appendix

A Basic Background of Parameter Efficient Transfer Learning

Since very little work has been done to study PETL in low-level vision, we re-implement the current state-of-the-art PETL methods in this work, such as VPT [9], Adapter [17], LoRA [21], AdaptFormer [7], SSF [24], and FacT [10]. In this part, we review these methods and provide detailed implementation details for reproduction. Fig. 7 gives an illustration of these baseline methods.

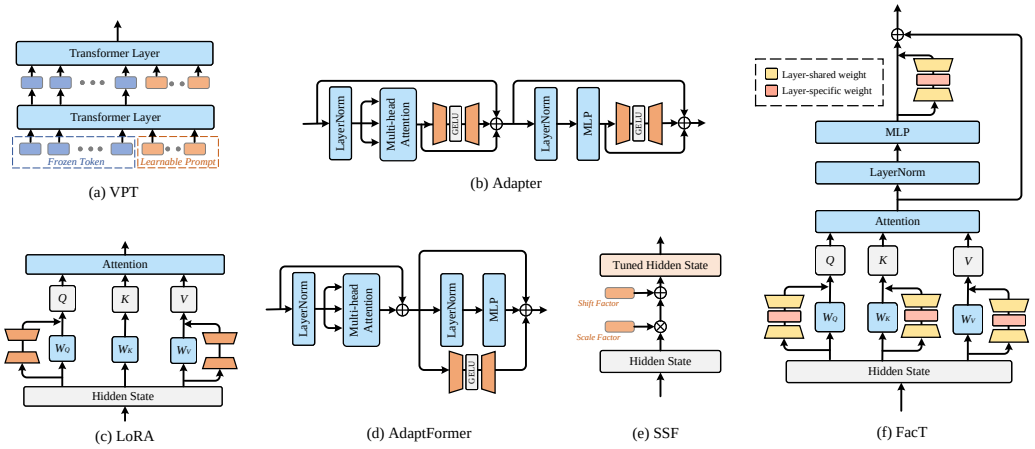


Figure 7: Illustrations of exsiting state-of-the-art PETL baselines.

- VPT [9], shown in Fig. 7(a), prepends learnable prompt tokens in the input of one transformer layer [35]. In [9], there are two versions of VPT, *i.e.*, $\text{VPT}_{\text{Shallow}}$ and VPT_{Deep}. To obtain a good performance, we use VPT_{Deep} which inserts new prompts at each transformer layer, as our default settings.
- Adapter [17], shown in Fig. 7(b), is a bottleneck structure with an intermediate GELU activation function. Following the vanilla Adapter design [17], we insert the Adapter both after the Multi-head Self-Attention (MSA) and Multi-Layer Perceptron (MLP).
- LoRA [21], shown in Fig. 7(c) use the multiplication of two low-rank to approximate the incremental matrices in projection layers of Query and Value.
- AdaptFormer [7], shown in Fig. 7(d), is similar in model architecture with Adapter, but has different insert position and form. In [7], the AdaptFormer is placed before the second LayerNorm layer and adopts a parallel insertion form.
- SSF [24], shown in Fig. 7(e), utilize learnable scale and shift factors to modulate frozen features. Based on the settings in [24], we place the SSF layer behind all the attention QKV projection, the LayerNorm, and the MLP layers.
- FacT [10], shown in Fig. 7(f), tensorises Vision Transformer [51] and introduces a low-rank approximation to the incremental matrix similar to LoRA. Different from LoRA, FacT sets the up-projection and down-projection to be shared across layers while setting the projection in the low-rank space to be layer-specific. There are two versions of FacT, namely FacT_{TT} and FacT_{TK} [10], we use FacT_{TT} in this work because of its good performance. Following [10], we introduce the FacT layer in attention QKV projection as well as MLP layers.

B Discussion with Prompt-based Methods

In this section, we briefly discuss the difference between our AdaptIR with other prompt-based methods. To the best of our knowledge, only the PromtGIP [52] shows the zero-shot ability when facing unseen degradations. And the other ProRes [53], PromptIR [5], and PromptRestorer [54] can only handle degradations which have been seen during training, which means they still need additional fine-tuning for task generalization. The advantages of our AdaptIR compared to these

https://doi.org/10.52202/079017-0432

Table 8: Comparison with prompt-based restoration methods.

Methods	Type	Fast Adaptation	Adaptation Cost	PSNR on denoising	PSNR on deraining
PromptGIP [52]	Prompt-based	Yes	zero-shot	26.22	25.46
ProRes [53]	Prompt-based	No	8x3090GPUs	Not open-source	Not open-source
PromptIR [5]	Prompt-based	No	7-days 8x3090GPUs	29.39	37.04
PromptRestorer [54]	Prompt-based	No	8x3090GPUs	Not open-source	Not open-source
Ours	PETL-based	Yes	8h 1x3090	29.70	37.81

prompt-based approaches are two-fold. As for efficiency, PromptIR, ProRes, and PromptRestorer all need full fine-tuning for adapting to new tasks, *e.g.*, PromptIR needs 7-day 8×3090 GPUs for full fine-tuning, while our AdaptIR needs only 8h 1×3090 GPU. As for performance, since these methods need to learn multiple degradations within one model, it is inevitable to suffer the problem of negative transfer, which impairs performance. We give a thorough comparison in Tab. 8.

C Discussion on Multi-task Restoration Paradigms.

In this section, we revisit existing paradigms for dealing with multi-task restoration problems, in which multiple degradations need to be handled. Let denote the number of degradation types, *i.e.*, downstream tasks as N, then we can summarize existing paradigm into the following three categories:

1. "N for N": training N task-specific models for N downstream tasks, such as Restormer, MPRNet.
2. "1 for N": training 1 all-in-one model for N downstream tasks, such as PromptIR [5], AirNet [4].
3. "(1+N) for N": using 1 task-shared pre-trained weights, and N task-specific lightweight modules.

Early image restoration techniques predominantly employed the first strategy, which trains N different models to handle multiple degradations. Although this strategy can handle multiple degradations, it usually requires training and storing N copies for each task. The recent all-in-one methods train one model for multiple degradations, although reducing the model copy to one to improve the efficiency, this approach usually suffers from performance degradation due to the multi-task learning difficulties and the negative transfer learning problem. In this work, the proposed AdaptIR is the first paradigm categorized in the third category, which trains a shared pre-trained backbone as well as N task-specific lightweight modules. This paradigm can be seen as a compromise between effectiveness and efficiency. However, given that the task-specific modules are very lightweight, we believe that the advantages of this paradigm outweigh the disadvantages.

D Further Explanation of the Heterogeneous Representation.

In this work, we pay attention to the learning of the *heterogeneous representation*. Here, we articulate it to make it more clear about this term. The heterogeneous representation in this paper represents the learning of discriminative features across different degradation types. The term representation here is instantiated as the Fourier curve in Fig. 1 in the main paper. Previous approaches tend to produce similar representations across various degradations. As common knowledge, restoring different degradations requires different representations, *e.g.*, SR needs a high-pass filter network while denoising needs low-pass. As a result, if the representation needed by current degradation matches the specific representation of the existing PETL method, it works. If not, it leads to unstable performance. To demonstrate the generality of the problem regarding the unstable performance and the homogeneous representation under different degradations, we provide more evidence in Fig. 9 and Fig. 10.

Table 9: Quantitative comparison for $\times 4$ image SR on PSNR(dB). We compare the #param when the performance is the same.

Method	#param	**Set5**	**Set14**	**BSDS 100**	**Urban 100**	**Manga 109**
RCAN [55]	15.6M	32.63	28.87	27.77	26.82	31.22
SAN [14]	15.9M	32.64	28.92	27.78	26.79	31.18
SwinIR [56]	11.9M	32.74	29.06	27.89	27.37	31.93
Full-ft	119M	32.66	29.03	27.82	27.31	31.64
pretrain	-	32.58	28.97	27.79	27.18	31.41
VPT [9]	884K	32.71	29.02	27.82	27.20	31.65
Adapter [17]	691K	32.70	29.03	27.82	27.21	31.68
LoRA [21]	995K	32.70	29.03	27.82	27.20	31.68
Adaptfor. [7]	677K	32.70	29.03	27.82	27.21	31.68
SSF [24]	373K	29.56	26.84	26.50	23.78	26.02
FacT [10]	537K	32.71	29.03	27.82	27.23	31.70
Ours	370K	32.71	29.04	27.82	27.22	31.70

Table 10: Quantitative comparison for color image denoising on PSNR(dB). We compare the #param when PSNR is the same.

Method	#param	**CBSD68**		**Urban100**	
		σ=30	σ=50	σ=30	σ=50
FFDNet [57]	0.8M	30.31	27.96	30.53	28.05
RDN [58]	15.6M	30.67	28.31	31.69	29.29
SwinIR [56]	11.9M	-	28.56	-	29.82
Full-ft	119M	30.75	28.39	32.01	29.72
Pretrain	-	30.73	28.36	31.94	29.68
VPT [9]	884K	30.74	28.37	31.97	29.68
Adapter [17]	691K	30.74	28.37	31.97	29.69
LoRA [21]	995K	30.75	28.38	31.98	29.70
AdaptFor. [7]	677K	30.73	28.37	31.97	29.68
SSF [24]	373K	30.07	27.64	29.79	27.01
FacT [10]	537K	30.75	28.38	31.98	29.70
Ours	515K	30.74	28.38	31.98	29.69

Table 11: Quantitative comparison for light rain streak removal on Rain100L dataset.

Method	#param	PSNR	SSIM
Full-ft	119M	42.14	0.9905
Pretrain	-	17.30	0.5488
VPT [9]	884K	41.74	0.9896
Adapter [17]	691K	41.95	0.9900
LoRA [21]	995K	41.89	0.9898
AdaptFor. [7]	677K	41.90	0.8992
FacT [10]	537K	40.61	0.9984
Ours	697K	42.09	0.9902

Table 12: Quantitative comparison for heavy rain streak removal on Rain100H dataset.

Method	#param	PSNR	SSIM
Full-ft	119M	32.23	0.9202
Pretrain	-	17.30	0.5488
VPT [9]	884K	30.87	0.8967
Adapter [17]	691K	30.99	0.8971
LoRA [21]	995K	31.16	0.9002
AdaptFor. [7]	677K	31.10	0.8992
FacT [10]	537K	29.70	0.8824
Ours	697K	31.23	0.9016

E More Experiment Results

E.1 Comparison on Single-degradation Tasks

Current PETL methods struggle to achieve stable performance due to homogeneous frequency characteristics and suffer performance degradation when not well aligned with the frequencies required by a specific degradation. Therefore, it needs multiple single-degradation tasks to obtain convincing evaluation results. Here, we give results of single-degradation tasks, including **image super-resolution** in Tab. 9, **color image denoising** in Tab. 10, **light deraining** in Tab. 11, **heavy deraining** in Tab. 12, and **low-light enhancement** in Tab. 13. It can be seen that our method maintains a stable best performance on most single-degradation tasks, with the second-best method varying across tasks. For example, the recent state-of-the-art methods FacT [10] obtains comparable performance with our AdaptIR, however, it suffers significant performance degradation on the subsequent light and heavy rain streak removal tasks. Another example can also be seen in LoRA [21], which performs the second best in heavy deraining but struggles with low-light image enhancement tasks. In contrast, our method is more stable, achieving consistent best performance across these single-degradation tasks. We also provide quantitative comparisons on the single-degradation restoration tasks in Fig. 12 and Fig. 13.

E.2 Additional Results on Hybrid Degradation with EDT

In order to demonstrate the generalizability of our AdaptIR, we choose IPT [11] and EDT [34] as pre-trained base models to evaluate the performance of different PETL methods. Due to the page limit, we mainly present the results of IPT in the main paper. Here, we give more experimental results with EDT. The results on second-order degradation LR4&JPEG30 with EDT are shown in Tab. 14. It can be seen that our method continues to achieve state-of-the-art performance by a significant margin. For example, our AdaptIR outperforms the second-best method FacT [10] by up to 0.15dB PSNR while using fewer tunable parameters. The results with EDT as the base model demonstrate the robustness of the proposed method.

https://doi.org/10.52202/079017-0432

Table 13: Quantitative comparison for low-light image enhancement with LOLv1 dataset.

Metric	Pretrain	UFormer [59]	RetinexNet [60]	FIDE [61]	VPT [9]	Adapter [17]	LoRA [46]	AdaptFor. [7]	FacT [10]	AdaptIR (ours)
#param	-	-	-	-	884K	691K	995K	677K	537K	697K
PSNR	7.64	16.36	16.77	18.27	19.28	19.22	18.94	19.40	19.06	19.46
SSIM	0.2547	0.771	0.560	0.665	0.7198	0.7293	0.7197	0.7352	0.7147	0.7441

Table 14: Quantitative comparison for second-order degradation with LR4&JPEG30 using EDT as pre-trained restoration models. The best results are **bolded**.

Method	#param	**Set5** PSNR/SSIM	**Set14** PSNR/SSIM	**BSDS100** PSNR/SSIM	**Urban100** PSNR/SSIM	**Manga109** PSNR/SSIM
Full-ft	11.6M	27.29/0.7800	25.58/0.6598	23.71/0.6768	25.11/0.6096	25.69/0.8043
Pretrain	-	25.08/0.6638	23.95/0.5847	21.51/0.5569	24.08/0.5580	22.60/0.6612
VPT [9]	311K	26.39/0.7367	24.84/0.6306	22.48/0.6101	24.75/0.5902	23.98/0.7365
Adapter [17]	168K	27.00/0.7698	25.36/0.6518	23.30/0.6566	25.01/0.6039	25.13/0.7848
LoRA [21]	155K	27.01/0.7694	25.36/0.6513	23.26/0.6551	25.00/0.6038	25.09/0.7837
AdaptFor. [7]	162K	27.03/0.7715	25.40/0.6533	23.32/0.6581	25.02/0.6048	25.19/0.7873
SSF [24]	117K	26.91/0.7664	25.33/0.6502	23.21/0.6519	24.98/0.6027	24.98/0.7801
FacT [10]	174K	27.01/0.7703	25.37/0.6521	23.30/0.6569	25.00/0.6041	25.14/0.7855
Ours	170K	**27.13/0.7739**	**25.44/0.6545**	**23.41/0.6620**	**25.04/0.6057**	**25.29/0.7903**

E.3 Results on More Degradations.

In this section, we further include another challenging degradation type, namely the real image denoising which is unseen during the pre-training phase and is the real-world degradation type, to further demonstrate the generalization of the proposed AdaptIR. We use the training and testing sets in the SIDD for this experiment. The experimental results are shown in Tab. 15. It can be seen that our AdaptIR maintains its superiority when transferring to real-world degradation. For instance, our method outperforms LoRA by 0.13dB PSNR. The above experimental results demonstrate the robustness of our methods.

Table 15: Results on real-world denoising tasks with SIDD datasets.

Methods	AdaptFor.	LoRA	Adapter	FacT	MoE	Ours
#param	677K	995K	691K	537K	667K	697K
PSNR	39.03	38.97	39.00	39.02	39.05	39.10

F Complexity Analysis

In this section, we theoretically analyze the parameter complexity of the proposed method. We omit the bias term as the corresponding parameter is small. Assume that the hidden dimension of the pre-trained restoration model [11, 34] is d and the dimension of intrinsic space in AdaptIR is d'. For the dimensional up and down operations, the number of parameters is $2dd'$. For the Local Interaction Module, assuming the convolution kernel size is K and the pre-defined rank of U, V is r, then the number of parameters of LIM with depth-separable design is $d'r + rK^2$. For the Frequency Affine Module, the total parameters of the amplitude and phase projection are $2d'$. For the Channel Gating Module, which contains the channel compression as well as the FFN, the number of parameters is $d' + 2d'\frac{d'}{a}$. As for the Adaptive Feature Ensemble, the compression convolution costs d' number of parameters, and $2d'\frac{d'}{b}$ is used in pooling FFN. Summing up the above terms gives $\frac{2(a+b)}{ab}d'^2 + (r + 4 + 2d)d' + rK^2$. In the implementation, we set $r = d'/2$, $a = 2$,$b = 8$, $K = 3$ and $d' = \frac{d}{\gamma}$. Therefore, the total parameter complexity of AdaptIR is $(\frac{2}{\gamma} + \frac{7}{4\gamma^2})d^2 + \frac{17}{2\gamma}d \sim \mathcal{O}(\frac{d^2}{\gamma})$.

G Feature Response Intensity Analysis.

To make it more clear how the proposed multi-branch structure works, we give the distribution of feature response intensity of three branches across various tasks, including SR, heavy deraining,

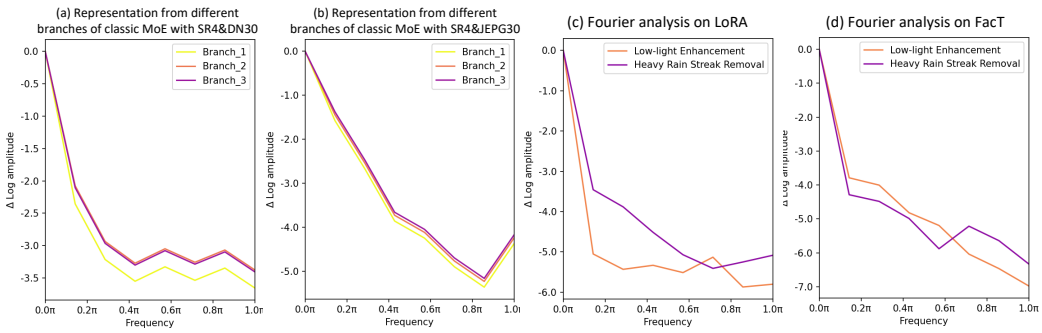


Figure 8: The frequency characteristic curves of features from three branches in the classic MoE with hybrid degradation of (a)SR4&DN30 and (b)SR4&JPEG30. (c)&(d) Fourier analysis on more current PETL methods, LoRA [21] and FacT [10], which shows significant representation homogeneity across tasks.

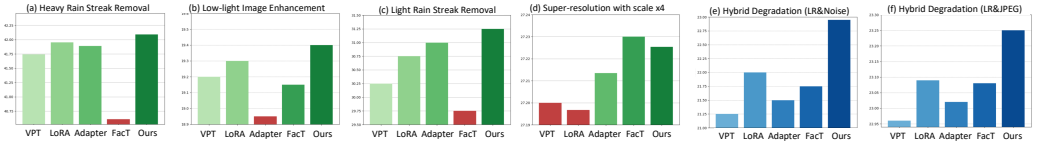


Figure 9: More evidence on the unstable performance of previous PETL methods across different single-degradation types, and the unfavorable performance under hybrid degradation.

light deraining, low-light image enhancement, and two hybrid degradations in Fig. 11. These figures indicate that our AdaptIR can adjust to different degradation types by enhancing or suppressing the outputs from different branches. Specifically, for the heavy&light deraining tasks, AdaptIR adaptively learns to enhance the low-frequency global features, i.e., the frequency affine module which is responsible for global spatial modeling has large values. This property ensures the removal of the high-frequency rainstreaks as well as the preservation of the global structure of the image. For SR tasks, AdaptIR adaptively enhances the restoration of local texture details by learning large output values from the local spatial modules. For the hybrid degradation task, AdaptIR shows it can distinguish between different hybrid degradations, i.e., three branches exhibit different patterns under two types of hybrid degradations. In short, each branch of AdaptIR can capture discriminative features under different degradations, indicating that our approach is degradation-aware. This ability guarantees robustness on single degradation and superior performance under hybrid degradation.

H Differences from Classic MoE

Although both our AdapIR and the classic MoE employ the multi-branch structure, however, our approach differs from the classic MoE in the following aspects. **Firstly**, the classic MoE uses the multi-branch structure to enhance the model capabilities, whereas our proposed heterogeneous MoE aims to capture heterogeneous representations across different restoration tasks. **Secondly**, despite using the multi-branch structure, the classical MoE still tends to capture homogeneous representations since each branch is the same, thus resulting in the sub-optimal results in Tab. 1. In contrast, each branch in our AdaptIR is designed orthogonally, thus ensuring the learning of orthogonal representation bases. **Thirdly**, classical MoE uses simple summation to fuse branches, which is degradation-agnostic, while our AdaptIR uses degradation-specific ensemble to learn the combination of orthogonal representation bases, facilitating heterogeneous representation across tasks.

In Fig. 8, we also give the frequency characteristics of the output features from different branches of the well-trained classical MoE. It can be seen that different branches still suffer from homogeneity despite the use of a multi-branch structure. In contrast, as shown in Fig. 4 in the main paper, our AdaptIR ensures that different branches capture different representations through the proposed orthogonal design, which promotes heterogeneous representations to achieve better performance.

https://doi.org/10.52202/079017-0432

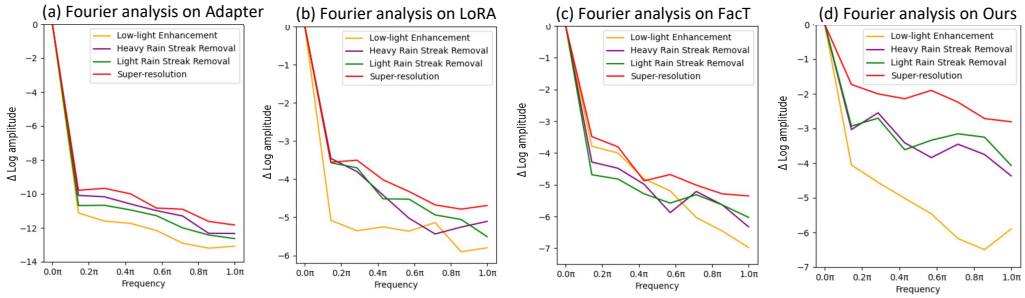


Figure 10: More evidence that shows previous PETL methods struggle to learn distinguishable features across different degradation types, i.e., homogeneous representation. In contrast, our AdaptIR can learn heterogeneous representations for different degradations.

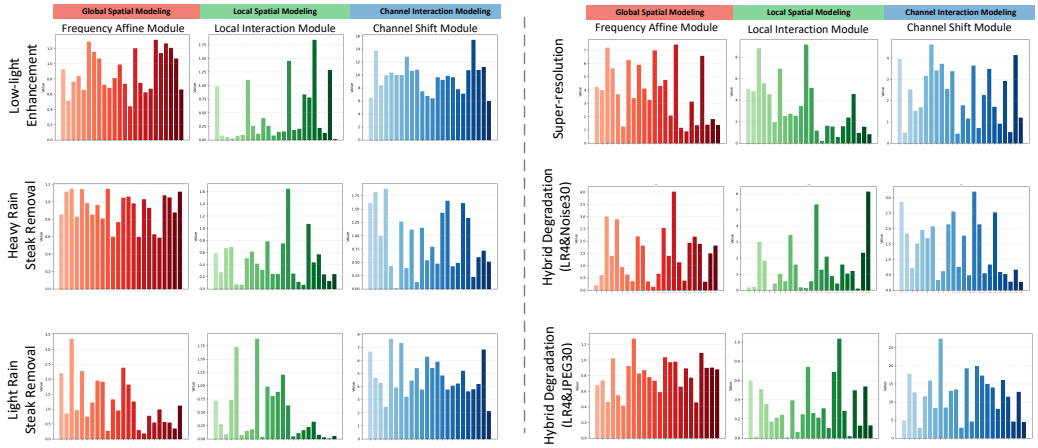


Figure 11: The distribution of feature response densities of the three branches across different tasks.

I More Evidence of Homogeneous Representation

In Fig. 1, we give the frequency characteristics of Adapter [8], and find its homogeneous representation when facing different degradations. To demonstrate the prevalence of homogeneous representations in current PETL methods, we provide the frequency characteristics curves of more PETL methods in Fig. 8. It can be seen that the current state-of-the-art PETL methods LoRA [21] and FacT [10] also exhibit homogeneity as Adapter, *i.e.*, the learned feature representations are similar even if they are for different degradations. In contrast, AdapIR utilizes the orthogonal multi-branch design to learn diverse representations, facilitating heterogeneous representations on different restoration tasks.

J Dataset Description

In this work, we evaluate different PETL methods on diverse image restoration tasks, which cover many training and testing datasets. To make the experimental setup more clear, we give a detailed description of datasets in Tab. 16.

K Limitations and Future Work

While AdaptIR appears as a competitive PETL alternative across various image restoration benchmarks, it can be further improved with task-specific module designs. For example, in the proposed AdaptIR, different tasks share the same structure, however, different restoration tasks have diverse model preferences. An intuitive solution might be to introduce degradation-aware dynamic networks. Moreover, although this work has covered multiple degradation types, some other degradations can also be explored in the future, *e.g.* blur and haze, to further demonstrate the generalization ability.

https://doi.org/10.52202/079017-0432

Table 16: Dataset description for various image restoration tasks.

Tasks	Type	Dataset	Num_samples
Super-resolution	train test	Div2K+Flicker2K Set5+Set14+BSDS100+Urban100+Manga109	800+2650 5+14+100+100+109
Denoise	train test	BSD400+WED BSD68+Urban100	400+4744 68+100
DerainL	train test	RainTrainL Rain100L	200 100
DerainH	train test	RainTrainH Rain100H	1800 100
Second-order Restor. (SR4&Dnoise30)	train test	Div2K+Flicker2K Set5+Set14+BSDS100+Urban100+Manga109	800+2650 5+14+100+100+109
Low-light Enhancement	train test	LOLv1-train-split LoLv1-test-split	485 15

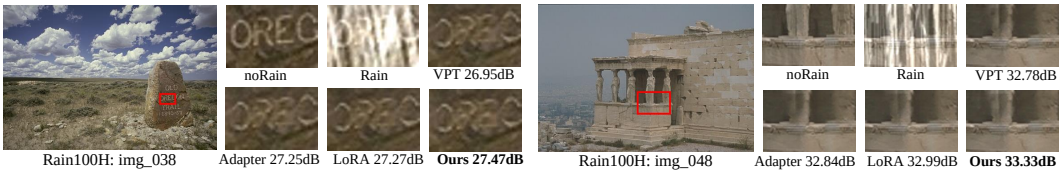


Figure 12: Visual comparison of heavy rain streak removal on samples from Rain100H [15] dataset.

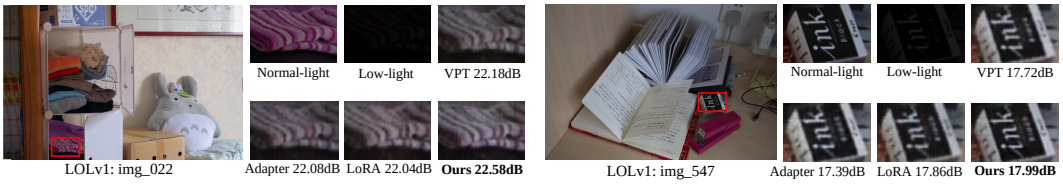


Figure 13: Visual comparison of low-light image enhancement on samples from LOLv1 [45] dataset.

L Broader Impact

Our AdaptIR holds significant promise for improving the quality and generalization of image restoration across various domains, such as medical imaging, historical document preservation, and digital media restoration. By enabling more accurate and reliable image restoration with reduced computational resources, AdaptIR can facilitate advancements in these fields, leading to better diagnostic tools, preservation of cultural heritage, and enhanced digital media quality. However, the enhanced capabilities of AdaptIR also present potential negative societal impacts, such as the risk of misuse in generating realistic fake images or deepfakes, which could be used for disinformation, creating fake profiles, or unauthorized surveillance, leading to privacy violations, security concerns, and ethical issues. To mitigate these risks, it is crucial to implement measures like gated releases of models, mechanisms for monitoring misuse and ensuring transparency in deployment and training processes, alongside continuous evaluation of the technology's impact.

https://doi.org/10.52202/079017-0432

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clearly outlined the main contributions of this paper in the abstract and introduction sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed in detail the limitation of this work in Appendix K.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions in the paper are either derived from the conclusions of previous research or validated through extensive experiments.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided a detailed description of our algorithm's process and implementation specifics in the paper, and we will release our code after review.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-0432

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will release our code after review, but we have already provided a detailed explanation of how to implement our algorithm and the specific implementation details in the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided detailed experimental details in both the paper and the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We have provided the performance fluctuation under different random seeds in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the computational resources we used for experiments in the implementation details section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have checked in every respect that this work conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have provided detailed discussion on possible both positive and negative societal impact in **??**.

    Guidelines:

https://doi.org/10.52202/079017-0432

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, the creators and original owners of assets used in this paper are properly credited, and the license and terms of use are explicitly mentioned and respected, as evidenced by thorough citations.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

https://doi.org/10.52202/079017-0432

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.