UHD-processer: Unified UHD Image Restoration with Progressive Frequency Learning and Degradation-aware Prompts

Supplementary Material

This supplementary document is organized as follows:

Sec. 1 outlines the motivation for using VAE and provides details of the VAE framework.

Sec. 2 elaborates on the details of the frequency-progressive learning strategy.

Sec. 3 provides the construction of the UHD all-in-one dataset and details of the experimental setup.

Sec. 4 provides additional visual results and experimental analysis.

Secs. 5 to 7 provide further discussions on related work, the limitations of this study, and its broader impacts.

1. More details on the VAE framework

1.1. The motivation for leveraging VAE as a resampling operator

Firstly, for UHD image processing, the high resolution presents a significant computational challenge compared to tasks involving regular-resolution images. To address this, a Downsampling-Restoration-Upsampling paradigm is commonly employed to reduce the computational load during the restoration process. However, the use of simple resampling operators often leads to considerable information loss, resulting in reconstructed outputs that lack fidelity. In contrast, VAE constructs a regularized latent space for image reconstruction, reducing information loss during the encoding process compared to other resampling operators. This approach facilitates more consistent and reliable reconstructions.

Secondly, for UHD all-in-one image restoration tasks, in addition to the computational burden imposed by high-resolution images, the model must also adapt to various degradation types. VAE handles this by encoding inputs with different degradation characteristics into a unified, compact latent space, effectively bridging the gap between these degradations. This simplifies the optimization process of the restoration network, improving overall performance.

However, directly applying VAE to UHD all-in-one image restoration tasks still presents three key challenges. First, while encoding into the latent space significantly reduces the computational complexity of the restoration mapping, the encoding and decoding processes of the VAE itself are not inherently efficient, and can still introduce substantial computational burdens. Second, although VAE-generated reconstructions maintain stronger semantic consistency, they often suffer from the loss of high-frequency details, which is unacceptable for low-level restoration tasks. Finally, while VAE's ability to map different degradations to the latent space helps reduce the domain gap in restoration mappings, VAE is trained on image reconstruction tasks using clean images. As a result, the domain gap during the encoding process still needs to be addressed.

1. To address the computational efficiency issue of VAE, we redesigned the VAE module by replacing conventional convolutions with depthwise separable convolutions and the original attention mechanism with a lightweight frequency-domain attention mechanism [31]. This results in a more lightweight version of VAE, termed Light-VAE, which is better suited for UHD-IR tasks.
2. To mitigate the high-frequency loss during the VAE encoding process, we introduced learnable adapter branches in both the encoder and decoder. We apply wavelet transform to the outputs at the corresponding scale of the encoder, extracting and injecting the mid-to-high frequency components into the decoder's adapter. This helps alleviate the loss of high-frequency details.
3. To address the domain gap in the VAE's encoding of different degradation types, we incorporate degradation-aware low-rank prompts into the encoder's adapter, as detailed in Section 3.3.1 of the main text. Further details on the adapter design are provided in the supplementary materials.

Here, we focus on the details of the adapter design that were not fully elaborated in the main text.

1.2. Details of adapter interactions

For the i_{th} Encoder, its input is $E_{i-1} \in \mathbb{R}^{h\times w\times c_1}$, and its output is $E_i \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times c_2}$. The corresponding adapter receives the output from the previous scale as the prior, $P^{(i-1)} \in \mathbb{R}^{h\times w\times 3}$, and applies wavelet transform to it as follows:

$$P^{(i)}_{LL}, P^{(i)}_{LH}, P^{(i)}_{HL}, P^{(i)}_{HH} = \mathrm{WT}(P^{(i-1)}), \tag{1}$$

Where, $P^{(i)}_{LL} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3}$ represents the low-frequency component of $P^{(i-1)}$, while $P^{(i)}_{High} = \{P^{(i)}_{LH}, P^{(i)}_{HL}, P^{(i)}_{HH}\} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3}$ correspond to the horizontal, vertical, and diagonal high-frequency components, respectively. Additionally, $P^{(0)}$ refers to the original resolution input image.

$$E'_i, P^{(i)'}_{LL}, P^{(i)'}_{High} = \mathcal{W}_{\mathrm{p}_{m\in\{1,2,3\}}}([E_i, P^{(i)}_{LL}, P^{(i)}_{High}, E^i_{pro}]), \tag{2}$$

where $\mathcal{W}_{\mathrm{p}}$ denotes pointwise convolution,E^i_{pro} denotes the encoder prompt, and E'_i serves as the output passed from the

adapter to the Encoder, which is then combined with the original output E_i through summation after applying a zero convolution. The component $P_{LL}^{(i)'}$ represents the low-frequency downsampling result from the adapter branch, acting as the prior input for the next-level adapter. Meanwhile, $P_{High}^{(i)'}$ corresponds to the high-frequency components, which are utilized as upsampled high-frequency complementary information for the corresponding adapter in the decoder.

In contrast to the encoding process, the decoder progressively increases the spatial scale of the features during the decoding process. For the i-th Decoder layer, the input is $D_{i-1} \in \mathbb{R}^{\frac{h}{2} \times \frac{w}{2} \times c_2}$, and the output is $D_i \in \mathbb{R}^{h \times w \times c_1}$. The corresponding adapter at this level receives the output from the previous level's adapter, $P_d^{(i-1)} \in \mathbb{R}^{\frac{h}{2} \times \frac{w}{2} \times 3}$, along with the high-frequency components from the Encoder adapter, $P_{High}^{(N+1-i)} \in \mathbb{R}^{\frac{h}{2} \times \frac{w}{2} \times 3 \times 3}$, where N denotes the total number of encoder or decoder blocks. The adapter first applies an inverse wavelet transform on these inputs to reconstruct $P^{(i)} \in \mathbb{R}^{h \times w \times 3}$, reintroducing high-frequency details into the spatial domain. This reconstructed output $P^{(i)}$ is then combined with the Decoder's output D_i through a zero convolution layer, initialized with zeros to ensure that the dimensions and characteristics of the features are preserved. This combination enables smooth integration of high-frequency details from the adapter with the structural features from the Decoder. The process can be formulated as follows:

$$D_i' = \mathcal{W}_{\text{zero}}(\text{IWT}(P_d^{(i-1)}, P_{High}^{(N+1-i)'}), D_{pro}^i) + D_i, \quad (3)$$

where $\mathcal{W}_{\text{zero}}$ is the zero convolution and IWT is the inverse wavelet transform, and D_{pro}^i is the decoder prompt.

2. More details about the frequency-progressive learning strategy

2.1. Metric for measuring the differences between tasks

As shown in Section 3.1, to measure the gap between tasks under low-frequency and high-frequency conditions, we introduce the metric z_i^t from [27], which is used to quantify the differences between pretraining and fine-tuning tasks. The specific definition is as follows:

$$\hat{z}_{ij}^s = \frac{1}{s_e - s_0} \sum_{s=s_0}^{s_e} z_{ij}^s = \frac{1}{s_e - s_0} \sum_{s=s_0}^{s_e} \frac{\mathcal{L}_i\left(\tau^i, \theta_s^i\right)}{\mathcal{L}_i\left(\tau^i, \theta\right)}, \quad (4)$$

Here, θ represents the parameters obtained after pre-training on task τ^j, and θ_s^i denotes the parameters after s steps of fine-tuning on task τ^i. Thus, z_{ij}^s measures the difference between two tasks by comparing the loss of the updated parameters after fine-tuning on a new task with the original parameters' loss. A larger discrepancy between tasks results in a faster decrease in loss after fine-tuning, yielding a smaller z_{ij}^s, and vice versa. To ensure stability, we take the average value from step s_0 to s_e.

Therefore, we train on one restoration task and then fine-tune on another, measuring $\hat{z}_{ij}^s$ to quantify the inter-task differences, with a smaller $\hat{z}_{ij}^s$ indicating a greater degree of difference. The results, as shown in Figure 1(d) of the main text, reveal that the gap between tasks in different frequency bands are smaller than those across the full frequency band, with the gap between low-frequency band being smaller than those between high-frequency band. This experimental observation supports the motivation behind our proposed frequency-progressive learning strategy.

2.2. Details of the Multiple Wavelet Transformation (MWT)

The formula for a single wavelet transform is given in Equation (1). By repeatedly applying wavelet transforms to the low-frequency components, we obtain multi-band decomposition results, which are expressed as Algorithm 1.

The corresponding inverse wavelet transform can be expressed as follows: For the inverse wavelet transform, it involves performing an inverse transform on each frequency band to reconstruct the input. Assuming the inverse wavelet transform is denoted as IWT, the formula can be expressed as Algorithm 2.

Algorithm 1 Multiple Wavelet Transform (MWT)

Input: F - original input image
Output: F_{MWT} - frequency bands after multiple wavelet transforms
Initialize $F_{LL}^{(0)} = F$ {Start with original input as the lowest frequency}
$F_{\text{MWT}} \leftarrow \emptyset$ {Initialize the list to store results}
for j = 1 to J **do**
    Perform $(F_{LL}^{(j)}, F_{\text{High}}^{(j)}) = \text{WT}(F_{LL}^{(j-1)})$ {Apply wavelet transform to get low and high frequencies}
    Append $(F_{LL}^{(j)}, F_{\text{High}}^{(j)})$ to F_{MWT}
end for
Return: F_{MWT} {Return the frequency bands}

3. Experimental Details

3.1. Datasets

We have developed a comprehensive benchmark for evaluating UHD all-in-one restoration performance, based on datasets from UHD-LL [11], UHD-blur [5], UHD-haze [28], UHD-rain [3], and UHD-haze [22]. Additionally, we constructed a denoising dataset, UHD-noise, using 4K images from [26] as the background.

The distributions of the training and testing sets for all datasets are shown in Tab. 1. For UHD-Rain, UHD-Snow,

Algorithm 2 Inverse Multiple Wavelet Transform (IMWT)

Input: F_{MWT} - frequency bands from MWT, J - number of wavelet transform cycles
Output: F_{rec} - reconstructed image
Initialize $F_{\text{rec}} \leftarrow F_{LL}^{(J)}$ {Start with the lowest frequency component from the last stage}
for j = J to 1 **do**
    $(F_{\text{High}}^{(j)}) \leftarrow F_{\text{MWT}}[j-1]$ {Get the high-frequency component for the current stage}
    $F_{\text{rec}} \leftarrow \text{IWT}(F_{\text{rec}}, F_{\text{High}}^{(j)})$ {Use the previous reconstructed low-frequency and current high-frequency to reconstruct}
end for
Return: F_{rec} {Return the reconstructed image}

Table 1. Dataset details and corresponding tasks.

Dataset	Training samples	Testing samples	Task
UHD-Snow	2,000	200	Desnowing
UHD-Blur	1,964	300	Deblurring
UHD-Rain	2,000	500	Deraining
UHD-LL	2,000	115	LLIE
UHD-Haze	2,290	231	Dehazing
UHD-Noise	2,000	500	Denoising

and UHD-Noise, we sample the training set according to different scenes, filtering out some repeated scenes while controlling the training set size to be comparable with other tasks. For all tasks, the entire testing set is used to better validate the generalization performance.

3.2. Implementation Details

The number of encoder and decoder layers is set to 3, and the number of cubic mixer blocks in the latent space restoration sub-network, from low frequency to high frequency, is set to [8, 6, 4, 2].

For the first stage, we train Light-VAE on the image reconstruction task. The initial learning rate is set to 5×10^{-4}, gradually reduced to 1×10^{-7} using cosine annealing [13]. The batch size is set to 16, and the images are randomly cropped to 256×256.

For the second stage, we train the UHD-processor on the image restoration task, keeping the parameters of Light-VAE frozen. We fine-tune the parameters of the adapter, prompt, and latent restoration network. The initial learning rate is set to 8×10^{-4}, gradually reduced to 1×10^{-7} using cosine annealing [13]. The batch size is set to 6, and the images are randomly cropped to 512×512.

3.3. Training procedure

In the first phase, Light-VAE is trained to perform the image reconstruction task using clean images. The clean input image is denoted as I_h, and the reconstruction result is denoted as I_{r1}.

The loss function of a vanilla VAE includes two main components: reconstruction loss and KL divergence loss. The reconstruction loss measures the difference between the decoder's output and the original input, encouraging consistency with the input [23]. The KL divergence loss regularizes the latent space, ensuring representations conform to the prior distribution. This regularization enhances structural coherence and continuity, leading to consistent reconstructions from similar inputs [30].

We follow this design and the reconstruction loss and KL divergence loss can be expressed as follows:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \frac{1}{N}\sum_{i=1}^{N} \|I_{\text{rl}}^{(i)} - I_{\text{h}}^{(i)}\|_1, \\ \mathcal{L}_{\text{KL}} &= D_{\text{KL}}(q(z|I)\|p(z)), \end{aligned} \tag{5}$$

where $q(z|I_h)$ is the approximate posterior distribution of the latent variable z given the input image I_h, and $p(z)$ is the prior distribution, typically chosen as a standard Gaussian distribution $\mathcal{N}(0, I)$. The KL divergence D_{KL} measures how much the distribution$q(z|I_h)$ diverges from the prior $p(z)$.

In addition, we further maintain the frequency domain consistency of the reconstruction results using the FFT loss, which can be denoted as:

$$\mathcal{L}_{\text{FFT}} = \frac{1}{N}\sum_{i=1}^{N} \|\text{FFT}(I_{\text{rl}}^{(i)}) - \text{FFT}(I_{\text{h}}^{(i)})\|_1. \tag{6}$$

In the second phase, the parameters of Light-VAE are frozen, and UHD-processor takes over the image restoration task. In this phase, the input degraded image is denoted as I_d, which corresponds to the clean image I_{gt}. The output of this restoration process is the restored image, denoted as I_{r2}.

We apply both the L1 loss and FFT loss between the restored image I_{r2} and the ground truth clean image I_{gt}, similarly to the first stage. The expressions for these losses remain the same, as follows:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \frac{1}{N}\sum_{i=1}^{N} \|I_{r2}^{(i)} - I_{gt}^{(i)}\|_1, \\ \mathcal{L}_{\text{FFT}} &= \frac{1}{N}\sum_{i=1}^{N} \|\text{FFT}(I_{r2}^{(i)}) - \text{FFT}(I_{gt}^{(i)})\|_1. \end{aligned} \tag{7}$$

4. More Experimental Results

4.1. More visual comparison results.

The visual results of the all-in-one restoration task for six types of degradation are shown in Figure 1. From the comparison results in the figure, it can be observed that our method not only achieves the highest PSNR but also produces the most visually pleasing results. Additionally, some

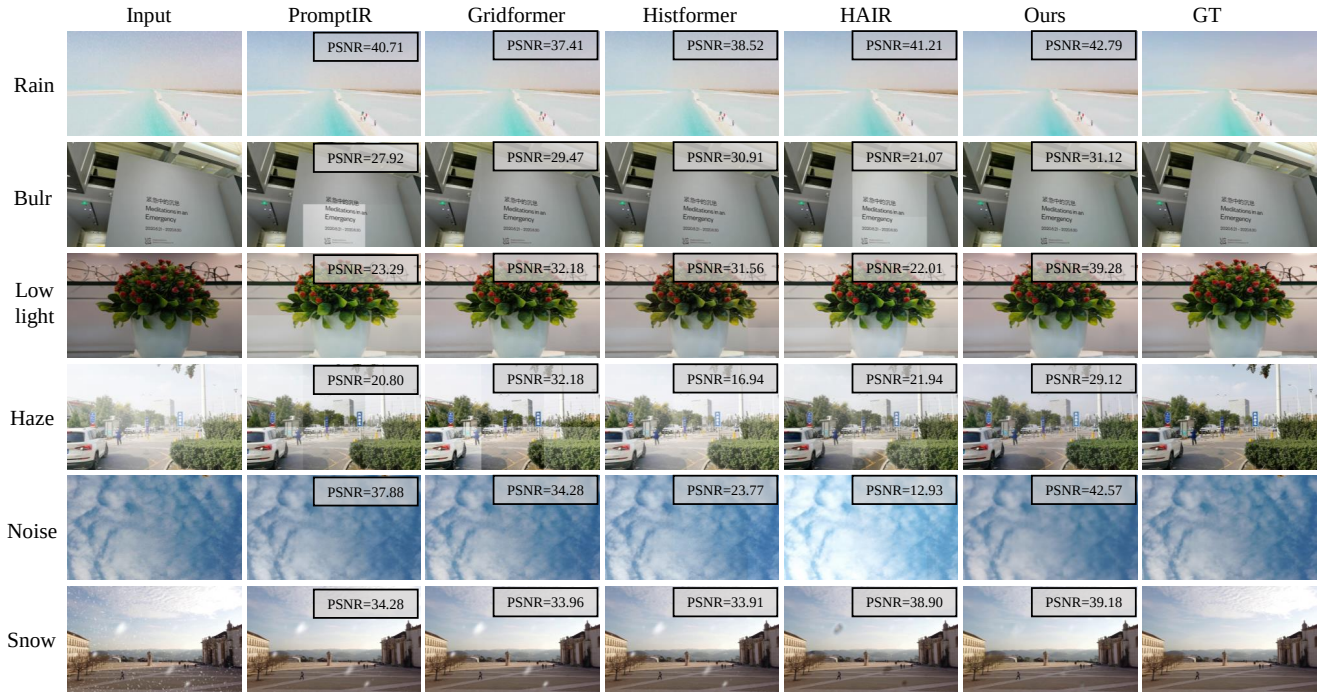

Figure 1. A comparison of visual results for six types of degradation removal with other state-of-the-art (SOTA) all-in-one methods.

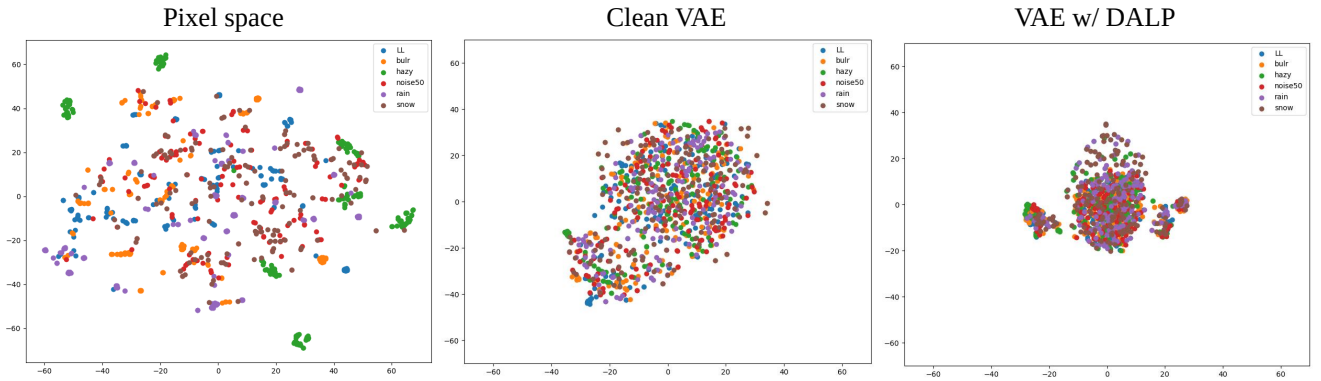

Figure 2. t-SNE clustering analysis of features from different degradation types. 'Clean VAE' refers to the VAE trained only on clean images. The results show that VAE with DALP effectively maps various degradations into a unified and compact latent space, significantly reducing the feature disparity between different types of degradations.

methods may suffer from mutual interference across different degradation types, leading to issues such as brightness enhancement in deblurring or denoising tasks, or artifacts in the snow removal task. In contrast, our method effectively avoids interference between tasks, resulting in consistent restoration outcomes.

4.2. The visualization analysis of Efficient Adaptive Prompt Learning

This section provides a visualization analysis of the Degradation-Aware Low Rank Prompt(DALR) during the encoding stage and the Degradation-Specific Frequency Selection Prompt(DFSP) during the decoding stage to validate their effectiveness.

4.2.1. Degradation-Aware Low Rank Prompt

We introduce VAE to transfer the restoration mapping process from pixel space to latent space. The compact latent space not only significantly reduces the feature scale but also narrows the gap between different degradations, thereby greatly reducing the computational load and optimization complexity for the latent-space restoration network. How-

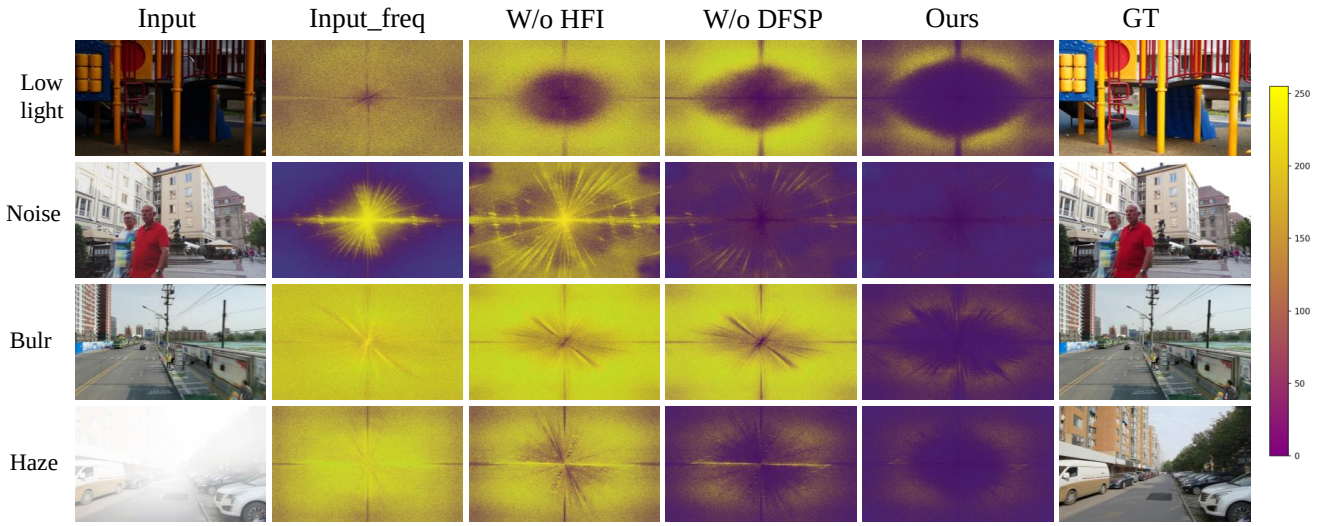

Figure 3. Ablation visualization analysis of the Degradation-Specific Frequency Selection Prompt (DFSP). We display Fourier magnitude spectrum residual maps between all results and the ground truth (GT). Input-freq refers to the magnitude spectrum residual between the Input and the GT. The color gradient from purple to yellow indicates an increasing difference with the GT magnitude spectrum. The results demonstrate that the integration of DFSP enables adaptive frequency selection and fusion based on degradation type, resulting in the most consistent spectral outcomes.

ever, since the first stage of VAE is trained for image reconstruction on clean images (Clean VAE), directly encoding degraded images leads to a domain gap, which impacts the encoding results. To address this issue, we introduce the Degradation-Aware Low Rank Prompt (DALR).

To validate the role of VAE and DALR, we perform a t-SNE feature clustering analysis. The results, as shown in Figure 2, indicate that using Clean VAE to encode different degradations results in more compact features compared to pixel space. By further introducing DALR, we overcome the domain gap in the VAE encoding process, achieving a more compact and unified latent space feature encoding. This significantly reduces the disparity between features of different degradations, allowing us to use a simple restoration network to handle the mapping process for multiple types of degradation.

4.2.2. Degradation-Specific Frequency Selection Prompt

Since the VAE encoding process often results in the loss of high-frequency information, leading to a loss of texture details in the final restoration, we first perform wavelet decomposition in the encoder stage's adapter. The high-frequency components are then injected into the decoder's adapter to facilitate high-frequency information injection(HFI).

However, due to the varying frequency bands affected by different types of degradation—such as low light and haze primarily influencing the low-frequency components, while noise and rain primarily affect the high-frequency components—simple High-Frequency Injection (HFI) cannot address these disparities effectively. Therefore, we introduce the Degradation-Specific Frequency Selection Prompt (DFSP) to enable adaptive frequency selection and fusion based on the type of degradation.

The visual analysis is shown in Figure 3, where we compare the spectral residuals between each result and the ground truth (GT). The results demonstrate that without HFI, the restoration results suffer significant high-frequency component loss. After incorporating HFI, this issue is partially alleviated, but due to the lack of degradation adaptation, the effectiveness is not guaranteed across all types of degradation. With the introduction of DFSP, the network can adaptively fuse frequency bands based on degradation information, minimizing the spectral difference with the GT across all degradation types.

4.3. More Ablation Experiments

4.3.1. Comparison experiments of different resampling methods

We compare our Light-VAE with other resampling methods (resamplers) in Tab. 2. In this comparison, "Laz3" refers to Lanczos3, "PS" to pixel shuffle [18], "AE" to autoencoder, and "LMAR" [24] refers to a learnable resampling operator. The term "one-stage" refers to end-to-end training of the resampler and enhancer, while "two-stage" refers to image reconstruction pretraining of the resampler, followed by freezing it during enhancer training for the restoration task. To ensure fairness, we kept the total parameters of all "resampler + enhancer" combinations consistent. The results demonstrate that our proposed method significantly outper-

forms other resamplers, indicating that the VAE encoded latent space is both compact and informative, effectively reducing computational costs while maintaining excellent performance.

4.3.2. Comparison experiment on Latent restoration network.

To validate the robustness of our proposed method, we conduct a comparative experiment using latent-space restoration networks built with different basic blocks. The experimental results, as shown in Tab. 3, demonstrate that our proposed UHD-processor is a versatile framework that consistently achieves good performance with various latent-space restoration networks. While a slight improvement is observed with the Restormer block, it introduces additional computational complexity and inference time. Therefore, balancing efficiency and performance, we choose the Cubic-Mixer as the basis for our latent-space restoration network.

4.3.3. Further discussions on sequential learning

The comparison results with sequential learning, shown in Tab. 4 of the main text, are further discussed here. Compared to sequential learning, our method has the following advantages:

First, we have a clear and defined learning order. While MIO-IR [10] establishes a rule for the sequential learning — starting with local degradation tasks like rain removal and denoising, followed by global degradation tasks like dehazing and low-light enhancement — this rule is effective but somewhat ambiguous when it comes to the internal ordering of tasks within both local and global degradation categories. Our experimental results indicate that this internal ordering significantly impacts the performance of sequential learning. This could be because the learning order is not only influenced by the type of degradation but also by specific dataset-related differences. Determining the learning order through experimental results is not ideal for practical applications. In contrast, our proposed frequency-progressive learning strategy (FDPL) has a task-independent learning order, and the relationships between frequency bands have clear physical significance. The training results from earlier stages effectively guide the learning process of subsequent stages, making our approach more generalizable.

Secondly, for all-in-one tasks, it is essential to avoid conflicts between tasks, but it is even more important to uncover the commonalities among them. Sequential learning methods, by gradually introducing tasks from fewer to more, initially isolate tasks to minimize interference. However, this isolation of tasks also leads to insufficient modeling of the shared characteristics between different degradations, with the model only beginning to capture the relationships between all degradations in the final stage. In contrast, our FDPL approach learns from all degradations across every frequency band, allowing for better capture of the relationships between degradations, and ultimately leading to improved restoration results.

Finally, for UHD image restoration tasks, UHD images contain more high-frequency details, making the restoration of high-frequency components significantly more challenging than for regular-resolution image restoration tasks. Therefore, the proposed FDPL, a frequency-progressive learning strategy from low to high frequencies, is a more suitable learning strategy for UHD all-in-one image restoration tasks.

Table 2. Ablation on resampling methods.FLOPs are computed based on an input size of 256 × 256.Inference time is tested on a 4K resolution using an RTX 3090.

Method	One Stage			Two Stage		
	Laz3	Bicubic	PS	AE	LMAR	Ours
PSNR	22.97	23.18	25.64	25.21	26.32	**29.23**
SSIM	0.763	0.746	0.832	0.821	0.834	**0.874**
Params	1.63M	1.62M	1.82M	**1.54M**	2.89M	1.60M
FLOPs	**3.22G**	3.24G	3.62G	3.86G	4.82G	4.17G
Runtime	1.04s	1.06s	1.12s	0.99s	1.23s	**0.98s**

Table 3. Comparison experiment on Latent restoration network. FLOPs are computed based on an input size of 256 × 256.Inference time is tested on a 4K resolution using an RTX 3090.

Method	PSNR ↑	SSIM ↑	Param ↓	FLOPs ↓	Runtime	FS
Restormer	**29.28**	0.869	4.17M	5.21G	1.21s	✓
NAFNet	29.16	0.872	<u>2.56M</u>	**4.12G**	1.01s	✓
Cubic-Mixer	<u>29.23</u>	**0.874**	**1.60M**	<u>4.17G</u>	**0.98s**	✓

5. Additional related work

5.1. Variational Autoencoder

The Variational Autoencoder (VAE) is a powerful generative model that has been widely recognized for its capability to learn compact representations from high-dimensional data. Unlike traditional autoencoders that map data to a single deterministic point in latent space, VAEs implement a probabilistic framework where latent variables are modeled as distributions rather than fixed points. This probabilistic element is central to VAEs, enabling them to produce diverse new data samples by sampling from these distributions. Essentially, VAEs compress input data into the parameters (mean and variance) of the latent distribution through an encoder and then regenerate the original data by sampling from this distribution via a decoder. VAEs have been successfully applied in a variety of fields, including image generation [2, 6, 9, 19], image compression [1, 4], and anomaly detection [15, 32].

In recent years, the compact and information-rich latent space encoded by Variational Autoencoders (VAEs) has increasingly drawn attention for its potential applications. Latent diffusion models [17] have leveraged this space to tran-

sition the diffusion process from pixel space to latent space, significantly boosting the efficiency of image generation. Old Photo Restoration (OPR) [20] employs a method of deep latent space translation to reduce the domain gap between synthetic degradation and real degradation, thereby enhancing the model's ability to generalize to authentic old photographs.

We aim to leverage the favorable properties of the VAE latent space to shift the restoration mapping process of UHD restoration tasks from the redundant pixel space to a compact latent space, thereby reducing computational complexity. At the same time, a unified latent space helps reduce the disparity between different degradations, which in turn eases the difficulty of the all-in-one UHD image restoration task.

5.2. Prompt learning

Prompt learning was initially explored as a method to integrate additional textual inputs, referred to as prompts, into pre-trained large language models to achieve specific outputs [25, 29]. As research progressed, the development of vision-prompt-based methods expanded, reducing reliance on textual information and fostering new techniques [8, 21]. In the field of image restoration, systems like ProRes [14] and PromptGIP [12] utilize additional input images or image pairs as prompts to specify the restoration tasks for networks. Similarly, PromptIR [16] and PIP [7] employ a classifier-based architecture to derive degradation details from images, using this information as representations for input-adaptive implicit prompting.

Inspired by these prompt-based learning methods, we integrate prompt learning into the VAE encoder and decoder. Based on the differences between the encoding and decoding processes, we design two types of prompts: a degradation-aware prompt and a frequency-selective prompt.

6. Limitations

Due to the limited number of real-world datasets available for UHD image restoration tasks, only the UHD-LL dataset [11] used in this paper is real, while the remaining datasets are synthetic. Therefore, constructing more real UHD degradation datasets is crucial for the field, as it will allow for a more accurate assessment of the proposed methods' performance in real-world scenarios. We will explore the creation of real UHD datasets in future work.

7. Broader Impacts

Due to the growing demand for high-resolution images across various industries, UHD image restoration tasks have become increasingly important. These tasks aim to restore images degraded by factors such as noise, blur, or low light, and they are essential in applications like medical imaging, satellite surveillance, and autonomous driving. Our proposed method for UHD all-in-one image restoration leverages a unified approach to address multiple degradation types, significantly improving restoration quality and computational efficiency. However, from a broader societal perspective, there are potential risks and considerations.

For instance, over-reliance on automated UHD restoration systems could lead to situations where restored images deviate from true real-world representations, especially in critical applications like medical imaging or security surveillance. In such scenarios, discrepancies between restored and actual images could potentially lead to misdiagnoses or incorrect security assessments. Furthermore, excessive reliance on automated restoration techniques could reduce the need for expert interpretation, which might compromise the quality of decision-making in high-stakes environments. Therefore, it is important to integrate human expertise into the restoration pipeline, ensuring that restored images are used with careful judgment. As this technology evolves, it will be crucial to strike a balance between automation and expert oversight to ensure that it benefits society while minimizing risks.

References

[1] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 6

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE international conference on computer vision*, pages 2745–2754, 2017. 6

[3] Hongming Chen, Xiang Chen, Chen Wu, Zhuoran Zheng, Jinshan Pan, and Xianping Fu. Towards ultra-high-definition image deraining: A benchmark and an efficient method, 2024. 2

[4] Ze Cui, Jing Wang, Shangyin Gao, Tiansheng Guo, Yihui Feng, and Bo Bai. Asymmetric gained deep image compression with continuous rate adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10532–10541, 2021. 6

[5] Senyou Deng, Wenqi Ren, Yanyang Yan, Tao Wang, Fenglong Song, and Xiaochun Cao. Multi-scale separable network for ultra-high-definition video deblurring. In *the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14030–14039, 2021. 2

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 6

[7] Hu Gao, Jing Yang, Ning Wang, Jingfan Yang, Ying Zhang, and Depeng Dang. Prompt-based all-in-one image restoration using cnns and transformer. *arXiv preprint arXiv:2309.03063*, 2023. 7

[8] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 7

[9] DP Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6

[10] Xiangtao Kong, Chao Dong, and Lei Zhang. Towards effective multiple-in-one image restoration: A sequential and prompt learning strategy. *arXiv preprint arXiv:2401.03379*, 2024. 6

[11] Chongyi Li, Chun-Le Guo, Man Zhou, Zhexin Liang, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Embedding fourier for ultra-high-definition low-light image enhancement. In *ICLR*, 2023. 2, 7

[12] 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. 7

[13] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[14] 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. 7

[15] Adrian Alan Pol, Victor Berger, Cecile Germain, Gianluca Cerminara, and Maurizio Pierini. Anomaly detection with conditional variational autoencoders. In *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, pages 1651–1657. IEEE, 2019. 6

[16] Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. Promptir: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 7

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 6

[18] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 5

[19] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 6

[20] Ziyu Wan, Bo Zhang, Dong Chen, Pan Zhang, Dong Chen, Fang Wen, and Jing Liao. Old photo restoration via deep latent space translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2071–2087, 2023. 7

[21] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, page 100047, 2023. 7

[22] Liyan Wang, Cong Wang, Jinshan Pan, Xiaofeng Liu, Weixiang Zhou, Xiaoran Sun, Wei Wang, and Zhixun Su. Ultra-high-definition image restoration: New benchmarks and a dual interaction prior-driven solution, 2024. 2

[23] Ronald Yu. A tutorial on vaes: From bayes' rule to lossless compression, 2020. 3

[24] Wei Yu, Jie Huang, Bing Li, Kaiwen Zheng, Qi Zhu, Man Zhou, and Feng Zhao. Empowering resampling operation for ultra-high-definition image enhancement with model-aware guidance. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25722–25731, 2024. 5

[25] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 7

[26] Kaihao Zhang, Dongxu Li, Wenhan Luo, Wenqi Ren, Bjorn Stenger, Wei Liu, Hongdong Li, and Yang Ming-Hsuan. Benchmarking ultra-high-definition image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[27] Wenlong Zhang, Xiaohui Li, SHI Guangyuan, Xiangyu Chen, Yu Qiao, Xiaoyun Zhang, Xiao-Ming Wu, and Chao Dong. Real-world image super-resolution as multi-task learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[28] Zhuoran Zheng, Wenqi Ren, Xiaochun Cao, Xiaobin Hu, Tao Wang, Fenglong Song, and Xiuyi Jia. Ultra-high-definition image dehazing via multi-guided bilateral learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16180–16189, 2021. 2

[29] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 7

[30] Lei Zhou, Chunlei Cai, Yue Gao, Sanbao Su, and Junmin Wu. Variational autoencoder for low bit-rate image compression. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2617–2620, 2018. 3

[31] Man Zhou, Jie Huang, Chun-Le Guo, and Chongyi Li. Fourmer: An efficient global modeling paradigm for image restoration. In *International conference on machine learning*, pages 42589–42601. PMLR, 2023. 1

[32] David Zimmerer, Simon AA Kohl, Jens Petersen, Fabian Isensee, and Klaus H Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection. *arXiv preprint arXiv:1812.05941*, 2018. 6