DreamUHD: Frequency Enhanced Variational Autoencoder for Ultra-High-Definition Image Restoration

Yidi Liu*, Dong Li*, Jie Xiao, Yuanfei Bao, Senyan Xu, Xueyang Fu†

School of Information Science and Technology and MoE Key Laboratory of
Brain-inspired Intelligent Perception and Cognition
University of Science and Technology of China, Hefei, 230026, China
{liuyidi2023, dongli6, ustchbxj, syxu}@mail.ustc.edu.cn, xyfu@ustc.edu.cn

Abstract

Existing ultra-high-definition (UHD) image restoration methods often struggle with consistency due to downsampling. We aim to address these challenges by leveraging the powerful latent space representation and reconstruction capabilities of Variational Autoencoders (VAE). However, applying VAE to UHD image restoration presents challenges: 1) High-performing VAEs have large parameter sizes, leading to significant carbon footprints; 2) The self-reconstruction property of VAE hinders bridging the domain gap between clean and degraded images; 3) Latent encoding in VAE can lose high-frequency information, compromising image detail. To overcome these challenges, we propose a frequency enhanced VAE UHD image restoration framework by integrating frequency priors. First, we design the Fourier-based lightweight frequency learning within the VAE to improve parameter efficiency. Then, we introduce a wavelet-based adapter that extracts multi-scale image information and employs frequency-aware adaptive modulation to bridge the domain gap by integrating degraded image data into the pre-trained VAE. Additionally, the adapter injects high-frequency information into the VAE decoder, enhancing detail in the restored images. In this way, our method effectively combines the powerful latent space representation with frequency priors to enhance UHD image restoration. Extensive experiments on various UHD image restoration tasks show that our method surpasses state-of-the-art methods both qualitatively and quantitatively.

Code — https://github.com/lyd-2022/dreamUHD

Introduction

In recent years, ultra-high-definition (UHD) imaging technology has made significant strides, driven by advancements in imaging sensors and displays. However, UHD images captured under challenging conditions, such as fast motion, haze, or low light, often suffer from quality degradation, severely impacting their visual quality and limiting their applicability in advanced visual tasks. Current learning-based image restoration algorithms struggle to handle UHD images effectively, and the increased number of pixels in these images challenges the efficiency of existing image processing methods. Therefore, the development of UHD-specific image restoration methods is crucial and has garnered significant attention.

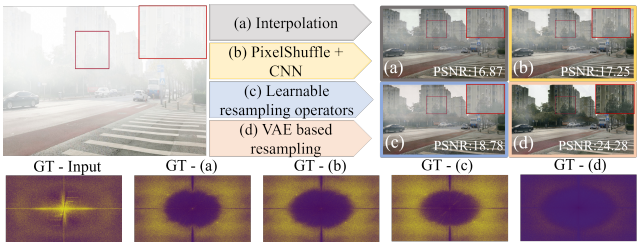

Figure 1: The impact of different resampling methods on UHD image restoration. We present the restoration results using Lanczos interpolation, PixelShuffle, learnable resampling operators (Yu et al. 2024), and our VAE-based resampling, along with the residuals between the results and ground truth in the Fourier amplitude spectrum. It is indicated that previous resampling methods not only lead to inconsistencies in the final restoration but also result in significant loss of high-frequency information.

Most existing UHD image restoration methods attempt to reduce computational costs by reducing resolution, typically following a "downsampling-enhancement-upsampling" paradigm (Zheng et al. 2021b,a). With the progress in deep learning, some works have replaced traditional interpolation-based upsampling and downsampling with learnable resampling operators, achieving remarkable results (Recasens et al. 2018; Sun and Chen 2020). However, this paradigm inevitably loses information during downsampling, and upsampling struggles to recover the details, leading to inconsistent restoration results, as illustrated in Figure 1. To compensate for the loss of details, some methods retain a full-resolution branch outside the resampling branch to supplement information (Wang et al. 2024b,a). However, these methods do not fundamentally address the limitations imposed by resampling, and the high computational burden of the full-resolution branch makes full-resolution inference on consumer-grade GPUs challenging. Therefore, how to ensure consistency as much as possible while reducing computational cost becomes the core problem of UHD im-

*Co-first authors contributed equally.

†Corresponding author.

age restoration. We believe that processing image features in an information-rich yet compact space offers a promising solution to this problem. Recent research in the generative domain has demonstrated the powerful latent space representation capabilities of the Variational Autoencoder (VAE), which also exhibits excellent image reconstruction abilities (Rombach et al. 2021). The VAE regularizes the output of the encoder by introducing a probabilistic distribution in the latent space. This regularization gives the latent representation strong structural integrity and continuity, enabling the VAE to generate robust compressed representations while capturing the underlying data distribution, thereby enhancing reconstruction consistency (Zhou et al. 2018; Yu 2020). These qualities make the VAE a promising approach for improving UHD image restoration. Therefore, we would like to leverage the pre-trained VAE as a resampler for UHD images, transferring image restoration to the latent space, which is much more compact than the pixel space, thereby improving consistency while reducing computational costs.

However, applying VAEs to UHD image restoration presents significant challenges: 1) High-performing VAEs have large parameter sizes, which would result in a substantial carbon footprint when applied to UHD image restoration. For example, in the latent diffusion model (Rombach et al. 2021), the parameter count and computational cost of the VAE are 83.6M and 445.3 GFlops, respectively, making full-resolution inference on consumer-grade GPUs impractical. 2) VAEs trained on clean images may encounter a domain gap when encoding degraded images, leading to mappings into an abnormal distribution that deviates from the clean image distribution. This issue arises primarily because the self-reconstruction property of VAEs tends to reconstruct the distribution of the input images. 3) The latent encoding in VAEs may lose high-frequency information, compromising the fidelity of image restoration and resulting in the loss of fine details in UHD images.

In this work, we propose a frequency enhanced VAE UHD image restoration framework (**FEVAE-UHD**), leveraging frequency-domain priors to address the challenges while harnessing the potential of the VAE. Specifically, to **mitigate the computational cost** of VAEs, we design the Fourier-prior lightweight frequency learning approach to improve VAE, thus proposing frequency enhanced VAE **(FE-VAE)**. The FE-VAE enhances parameter efficiency by integrating the Fourier domain's global perceptual capability and the ability of the Fourier spectrum to decompose image degradation, significantly reducing the parameter count while maintaining comparable performance. To **address the domain gap and high-frequency information loss**, we design a wavelet-based adapter that performs multiple functions through a carefully designed mechanism. This adapter extracts multi-scale image information via progressive wavelet transforms and integrates degraded features into the VAE encoder at corresponding scales. During this process, to eliminate the domain gap between the degraded images and the VAE trained on clean images, the adapter includes a frequency-aware adaptive modulation module that performs weighted modulation in both spatial and frequency domains, adaptively injecting degraded information into the VAE. Finally, the adapter selectively injects high-frequency information from the wavelet transform into the VAE decoder to enhance the high-frequency details in the restored images. In this way, our method overcomes the primary challenges of applying VAEs to UHD image restoration, including parameter efficiency, domain gap, and high-frequency information loss, thereby effectively combining the powerful latent space representation of VAEs with the enhancement provided by frequency priors.

In summary, our contributions are as follows:

- We propose a VAE-based UHD image restoration framework, FEVAE-UHD, which integrates frequency priors to utilize the powerful latent space representation of the VAE to improve consistency. To the best of our knowledge, this is the first work to introduce VAE into UHD image restoration.
- We design a Fourier-based frequency-enhanced VAE, FE-VAE, which significantly reduces computational and parameter costs while maintaining comparable representation capability.
- We develop a wavelet-based adapter to supplement the high-frequency details required for image restoration, while combining spatial and frequency information to establish a bridge between the frozen VAE and the degraded information, thereby mitigating the domain gap.

Extensive experiments demonstrate the effectiveness of our method, achieving state-of-the-art results in various tasks: UHD low-light image enhancement, UHD image deblurring, UHD image dehazing, UHD low-light image enhancement, and UHD moiré pattern removal.

Related Works

Variational Autoencoder. Recent progress in deep learning (Rombach et al. 2021; Li et al. 2023b; Duan et al. 2024; Zhu et al. 2024a; Li et al. 2025) include the Variational Autoencoder (VAE) (Kingma 2013), a generative model that learns a low-dimensional data representation in latent space. Unlike traditional AE with fixed latent variables, the VAE models them as probability distributions, enabling diverse data generation. It compresses input data into the mean and variance of the latent distribution via an encoder and reconstructs it by sampling from this distribution through a decoder. VAEs are widely used in tasks such as image generation (Rombach et al. 2021), compression (Duan et al. 2024), and anomaly detection (Chen et al. 2019).

Ultra-High-Definition Restoration. Image restoration tasks have received significant attention in recent years (Zamir et al. 2022; Chen et al. 2022; Kong et al. 2023a; Chen et al. 2023; Li et al. 2023c; Zhu et al. 2024b; Ge et al. 2024; Peng et al. 2024; Xiao et al. 2024).As the demand for UHD image processing grows, several approaches have emerged to recover clear UHD images through various network architectures. These include bilateral learning for image dehazing (Zheng et al. 2021a), multi-scale separable-patch networks for video deblurring (Deng et al. 2021), and Fourier embedding networks for low-light image enhancement (Li et al. 2023a). Recently, generalized architectures for UHD

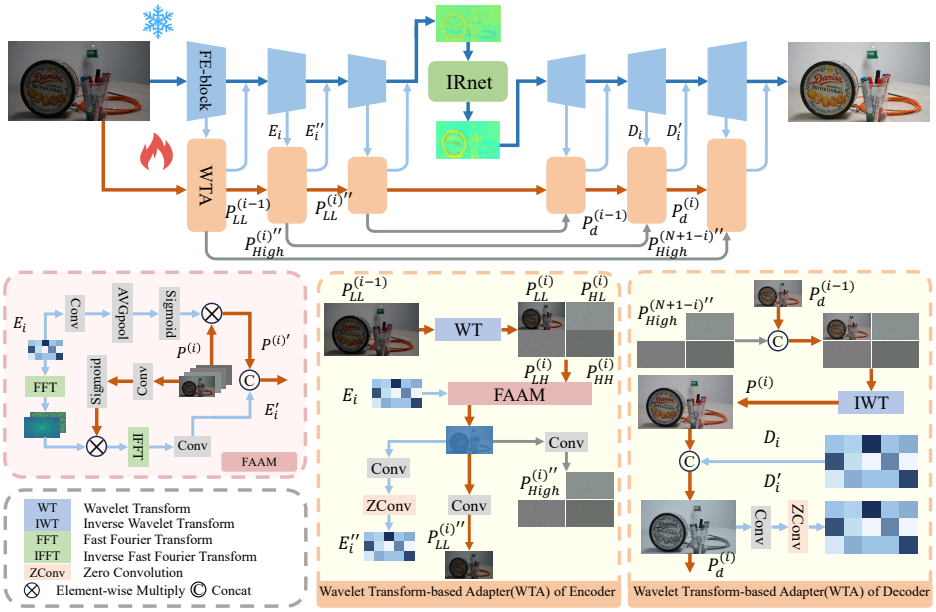

Figure 2: An overview of the proposed framework FEVAE-UHD. It consists of three main components: the frozen FE-VAE, the Wavelet Transform-based Adapter, and the arbitrary restoration network IRNet. In this stage, FEVAE-UHD training is based on the image restoration task.

restoration have rapidly developed, targeting multiple problems through unified structures. For example, (Wang et al. 2024a) introduces a correlation matching mechanism between high- and low-resolution branches, while (Wang et al. 2024c) incorporates gradient and normal priors. However, the resampling operations in these methods do not optimally balance computational efficiency and information retention.

Methodology

The framework of our FEVAE-UHD is illustrated in Fig.2. Our approach follows a two-stage process: in the first stage, we train the FE-VAE using an image reconstruction task, as illustrated in Fig.3. In the second stage, we freeze the VAE's weights and then perform image restoration based on the VAE's latent code. For the input degraded image $I \in \mathbb{R}^{H\times W\times 3}$, the VAE transforms it into a latent representation $L_{\text{in}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$, performs restoration in the latent space, and decodes the restored latent representation $L_{\text{out}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$ to generate the final image. To alleviate the loss of high-frequency components and the domain gap issues in the pre-trained VAE, we introduce a wavelet-based adapter (WTA) branch in parallel with the VAE backbone.

Frequency Enhanced VAE

Architecture The architecture of the Frequency Enhanced VAE (FE-VAE) is illustrated in Fig.3. To enhance efficiency while preserving representational capability, we integrate Fourier learning into FE-VAE. The FE-VAE framework comprises two key components: 1) Space-Frequency Adaptive Decomposition (**SFAD**), which employs channel-wise Fourier sensing to decompose input features into spatial and frequency branches, and 2) Frequency-Aware Feature Extraction (**FAFE**), which incorporates frequency learning for feature extraction.

We aim to leverage the efficient computation in the frequency space to lightweight the feature extraction process, while minimizing the sacrifice of the model's spatial perception ability. One straightforward approach is to directly split the input features along the channel dimension, with one part performing traditional spatial-domain computation and the other part performing efficient frequency-domain computation. However, this direct division is simplistic and lacks global channel awareness as well as an adaptive selection process.

To refine the channel split and adaptively select features for frequency or spatial domain processing, we propose a more nuanced approach. For an input feature $F \in \mathbb{R}^{h\times w\times c}$, instead of directly dividing the channels, we first apply a channel-wise Fourier transform to establish global channel perception, extracting the amplitude A_{c} and phase P_{c}. Since the Fourier spectrum captures global spatial information, a pointwise convolution effectively aggregates this information. Noticing that degradation primarily affects the magnitude, we apply convolution only to the magnitude, leaving the phase unchanged. This reduces computational complexity. The process is formulated as follows:

$$\begin{aligned} A_{\text{c}}, P_{\text{c}} &= \mathcal{F}_{\text{c}}(F), \\ F' &= \mathcal{F}_{\text{c}}^{-1}(\mathcal{W}_{\text{p}}(A_{\text{c}}), P_{\text{c}}), \end{aligned} \tag{1}$$

where $\mathcal{W}_{\text{p}}$ denotes the point-wise convolution, $\mathcal{F}_{\text{c}}(\cdot)$ and $\mathcal{F}_{\text{c}}^{-1}(\cdot)$ denote the channel dimensional Fourier transform

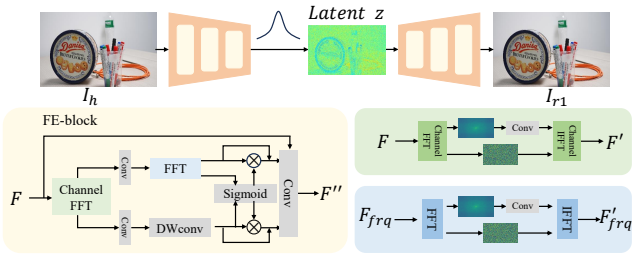

Figure 3: Frequency-Enhanced VAE. In the first stage, we train it using a reconstruction task. It consists of a stack of FE-blocks, each containing two key components: Space-Frequency Adaptive Decomposition (SFAD) and Frequency-Aware Feature Extraction (FAFE).

and the corresponding inverse transform, respectively.

This adaptive splitting enables the model to efficiently allocate features to the most appropriate processing domain and allows us to capture global information across channels. Following this, we use two point-wise convolutions to generate $F_{\text{frq}} \in \mathbb{R}^{h \times w \times \frac{c}{2}}$ and $F_{\text{sp}} \in \mathbb{R}^{h \times w \times \frac{c}{2}}$:

$$F_{\text{frq}} = \mathcal{W}_{\text{p}_1}(F'); F_{\text{sp}} = \mathcal{W}_{\text{p}_2}(F'). \tag{2}$$

In frequency domain learning, we first apply a Fourier transform to F_{frq} along the spatial dimensions, resulting in $F_{\text{four}} \in \mathbb{R}^{\frac{h+1}{2} \times \frac{w+1}{2} \times c}$. As before, we perform pointwise convolution only on the transformed amplitude spectrum. For spatial learning, we utilize a standard ResNet block but replace the conventional convolution with depth-separable convolution, which helps reduce both the number of parameters and computational costs. The process is formulated as follows:

$$\begin{aligned} A, P &= \mathcal{F}_{\text{hw}}(F_{\text{frq}}), \\ F'_{\text{frq}} &= \mathcal{F}^{-1}_{\text{hw}}(\mathcal{W}_{\text{p}}(A), P), \\ F'_{\text{sp}} &= \text{Resblock}(F_{\text{sp}}), \end{aligned} \tag{3}$$

where $\mathcal{F}_{\text{hw}}(\cdot)$ and $\mathcal{F}^{-1}_{\text{hw}}(\cdot)$ denote the spatial dimensional Fourier transform and the corresponding inverse transform.

After obtaining the features from both branches, we propose a Space-Frequency Interaction Module (**SFIM**) to ensure effective spatial-frequency interaction between the two branches. This module facilitates the effective exchange of information between the spatial and frequency domains, ultimately producing the final output. The process can be formulated as follows:

$$\begin{aligned} F''_{\text{frq}} &= \mathcal{S}(F'_{\text{sp}}) \odot F'_{\text{frq}} \\ F''_{\text{sp}} &= \mathcal{S}(F'_{\text{frq}}) \odot F'_{\text{sp}}, \\ F'' &= \mathcal{W}_{\text{p}}(\text{cat}[F''_{\text{frq}}, F''_{\text{sp}}]), \end{aligned} \tag{4}$$

where $\mathcal{S}$ denotes the sigmoid activation function, and $\text{cat}[\cdot]$ denotes channel concatenation.

Optimization In the first phase, FEL-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 (Yu 2020). The KL divergence loss regularizes the latent space, ensuring representations conform to the prior distribution. This regularization enhances structural coherence and continuity, resulting in consistent reconstructions for similar inputs (Zhou et al. 2018).

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:

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

Frequency Enhanced VAE UHD Image Restoration

Wavelet Transform-based Adapter After the comprehensive lightweighting of the VAE, we need to address the remaining two issues: reducing high-frequency loss during the encoding process and mitigating the domain gap of the Encoder for degraded domains. Therefore, we propose a Wavelet Transform-based Adapter(**WTA**).

This adapter allows efficient fine-tuning, keeping the pretrained VAEs frozen during restoration. By updating only the adapter's parameters, VAEs adapt to unknown degraded domains without altering the original VAE.

To preserve high-frequency details, we implement frequency-domain replay encoding and high-frequency complementary decoding. The encoding process mitigates high-frequency loss, while the decoding phase restores and enhances high-frequency details using the subbands generated by the adapter. The process is defined as follows:

Frequency Replay Encoding. The spatial scale of the features decreases progressively during the Encoder's encoding process. For the i-th Encoder layer, the input is $E_{i-1} \in \mathbb{R}^{h \times w \times c_1}$, and the output is $E_i \in \mathbb{R}^{\frac{h}{2} \times \frac{w}{2} \times c_2}$. The corresponding adapter uses the output of the previous scale as the prior, denoted as $P_{LL}^{(i-1)} \in \mathbb{R}^{h \times w \times 3}$. The wavelet transform is then applied to $P_{LL}^{(i-1)}$ as follows:

$$P_{LL}^{(i)}, P_{LH}^{(i)}, P_{HL}^{(i)}, P_{HH}^{(i)} = \text{WT}(P_{LL}^{(i-1)}), \tag{7}$$

where $P^{(i)} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3}$ represents low-frequency component of $P_{LL}^{(i-1)}$, while $P_{LH}^{(i)}, P_{HL}^{(i)}, P_{HH}^{(i)} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3}$ represent its horizontal, vertical, and diagonal high-frequency components. And $P_{LL}^{(0)}$ i.e., the original resolution input image.

To fuse E_i with the four subbands obtained from the wavelet transform $P^{(i)} \in \mathbb{R}^{\frac{h}{2}\times\frac{w}{2}\times 3\times 4}$, we design a frequency-aware adaptive modulation module(**FAAM**). This module first learns four filter kernels, each corresponding to one of the four subbands in $P^{(i)}$, to modulate the spectrum of E_i. This modulation captures the multi-frequency representation by incorporating information from the respective subband. After modulation, we apply a pointwise convolution to the modulated outputs to obtain the fused representation, denoted as E_i'. The process can be formulated as follows:

$$\begin{aligned} \text{Filter}_s &= \text{cat}[\mathcal{W}_{\text{c}}(\mathcal{S}(P^{(i)}))], \\ E_i' &= \mathcal{W}_{\text{p}}(\mathcal{F}_{\text{hw}}^{-1}(\text{Filter}_s \odot \mathcal{F}_{\text{hw}}(E_i))), \end{aligned} \tag{8}$$

For the subbands $P^{(i)}$, we use weights generated from channel pooling of E_i to modulate $P^{(i)}$, facilitating the interaction between features from different frequency bands and resulting in the modulated subbands $P^{(i)'}$. We then concatenate E_i' and $P^{(i)'}$ and apply three independent pointwise convolutions. This process yields the refined encoding representation E_i'', the low-frequency components $P_{LL}^{(i)''}$, and the high-frequency components $P_{High\in\{LH,HL,HH\}}^{(i)''}$, respectively. The process can be formulated as follows:

$$\begin{aligned} P^{(i)'} &= \mathcal{S}(\mathcal{A}_{\text{pool}}(E_i')) \odot P^{(i)} + P^{(i)}, \\ E_i'', P_{LL}^{(i)''}, P_{High}^{(i)''} &= \mathcal{W}_{\text{p}_{m\in\{1,2,3\}}}(\text{cat}[E_i', P^{(i)'}]), \end{aligned} \tag{9}$$

where $\mathcal{A}_{\text{pool}}$ denotes average pooling, E_i'' serves as the output passed from the adapter to the Encoder and is combined with the original output E_i through summation after applying a zero convolution. The component $P_{LL}^{(i)''}$ is used as the low-frequency downsampling result of the adapter branch, acting as the prior input for the next-level adapter. Meanwhile, $P_{High}^{(i)''}$ represents the high-frequency components, which are utilized as upsampled high-frequency supplementary information for the corresponding adapter in the decoder.

In this process, each downsampled output is fully fused with frequency information from the previous scale, introducing an effective frequency replay mechanism. This mechanism is crucial in mitigating the loss of frequency components during encoding.

High-frequency Injection Decoding. In contrast to the encoding process, the decoder progressively enlarges the spatial scale of the features during decoding. 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 number of encoders or decoders stages. The adapter first performs a wavelet inverse transform on these inputs to reconstruct $P_d^{(i)} \in \mathbb{R}^{h\times w\times 3}$, reintroducing high-frequency details into the spatial domain. This reconstructed output $P_d^{(i)}$ is then combined with the Decoder's output D_i through a zero convolution layer, which is initialized with zeros to ensure the dimensions and characteristics of the features are preserved. This combination allows for a smooth integration of the high-frequency details from the adapter and the structural features from the Decoder, the process can be formulated as follows:

$$D_i' = \mathcal{W}_{\text{zero}}(\text{IWT}(P_d^{(i-1)}, P_{High}^{(N+1-i)''})) + D_i, \tag{10}$$

where $\mathcal{W}_{\text{zero}}$ is the zero convolution and IWT is the inverse wavelet transform.

Image Restoration Network in Latent Space When the UHD input image is mapped from pixel space to latent space, its spatial size is reduced, and the feature distances between clean and degraded images become closer. This facilitates the use of a simple image restoration network in the latent space. Since our VAE already provides an informative representation, no additional complex designs are necessary to achieve high-quality results. As a result, our latent space restoration network(**IRNet**) utilizes only a few layers of the efficient Cubic-Mixer block (Zheng and Jia 2022). Further experiments with different latent space restoration networks are discussed in the ablation experiments section.

Optimization In the second phase, the parameters of FEL-VAE are frozen, and FEVAE-UHD 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:

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \frac{1}{N}\sum_{i=1}^{N} \|I_{r2}^{(i)} - I_{gt}^{(i)}\|_1, \\ \mathcal{L}_{\text{FFT}} &= \frac{1}{N}\sum_{i=1}^{N} \|\text{FFT}(I_{r2}^{(i)}) - \text{FFT}(I_{gt}^{(i)})\|_1. \end{aligned} \tag{11}$$

Experiments

We evaluate **FEVAE-UHD** on benchmarks for 4 UHD image restoration tasks: **(a)** low-light image enhancement, **(b)** image dehazing, **(c)** image deblurring, and **(d)** Image Demoiréing.

Main Results

Image Dehazing Results Tab.1 presents the quantitative dehazing results on UHD-Haze. Our FEVAE-UHD outperforms MB-TaylorFormer(Qiu et al. 2023) by over 3 dB in

Method	Venue	PSNR ↑	SSIM ↑	Param ↓	FS
UHD	ICCV'21	18.043	0.8113	34.5M	✔
Restormer	CVPR'22	12.718	0.6930	26.1M	✗
Uformer	CVPR'22	19.828	0.7374	20.6M	✗
DehazeFormer	TIP'23	15.372	0.7245	2.5M	✗
MB-TaylorFormer	ICCV'23	20.994	0.9194	2.7M	✗
UHDformer	AAAI'24	22.586	0.9427	**0.3393M**	✔
UHDDIP	arxiv'24	22.147	0.9418	0.81M	✔
Ours	-	**24.357**	**0.9454**	1.215M	✔

Table 1: Quantitative results of image dehazing.

Method	Venue	PSNR ↑	SSIM ↑	Param ↓	FS
MIMO-Unet++	ICCV'21	25.025	0.7517	16.1M	✔
Restormer	CVPR'22	25.210	0.7522	26.1M	✗
Uformer	CVPR'22	25.267	0.7515	20.6M	✗
Stripformer	ECCV'22	25.052	0.7501	19.7M	✗
FFTformer	CVPR'23	25.409	0.7571	16.6M	✗
UHDformer	AAAI'24	28.821	0.8440	**0.3393M**	✔
UHDDIP	arxiv'24	28.283	0.8452	0.81M	✔
Ours	-	**29.338**	**0.8523**	1.456M	✔

Table 2: Quantitative results of image deblurring.

PSNR while significantly reducing parameters. Compared to UHDformer and UHDDIP, FEVAE-UHD also shows notable gains in both PSNR and SSIM metrics. Fig. 4 illustrates that FEVAE-UHD produces clearer results, whereas other methods leave significant haze.

Image Deblurring Results We evaluate UHD image deblurring on the UHD-Blur dataset. Tab. 2 shows that FEVAE-UHD outperforms state-of-the-art methods, including a nearly 4dB gain over FFTformer (Kong et al. 2023b). Significant improvements are also noted over UHDformer and UHDDIP. Figure 5 visually demonstrates our method's superior ability to restore details like wall text and license plate numbers.

Low-Light Image Enhancement Results We evaluate low-light enhancement on the UHD-LL dataset. Tab.3 shows that FEVAE-UHD outperforms state-of-the-art methods, including Wavemamba (Zou et al. 2024), with fewer parameters. Fig.6 visually demonstrates that FEVAE-UHD produces more natural colors than other methods.

Image Demoiréing Results We conduct image demoiréing experiments on the UHDM dataset, as shown in Tab. 4. Our method outperforms the state-of-the-art ESDNet and its larger version, ESDNet-L, in objective metrics, while using fewer parameters. UHDformer and UHDDIP exhibit poor performance, highlighting the strong generalization capability of our approach.

Ablation Study

We use the UHD-Bulr dataset to conduct the ablation study on the main designs of FEVAE-UHD.

Method	Venue	PSNR ↑	SSIM ↑	Param ↓	FS
Restormer	CVPR'22	21.536	0.8437	26.1M	✗
Uformer	CVPR'22	21.303	0.8233	20.6M	✗
LLformer	AAAI'23	24.065	0.8580	13.2M	✗
UHDFour	ICLR'23	26.226	0.9000	17.5M	✔
UHDformer	AAAI'24	27.113	0.9271	**0.3393M**	✔
LMAR	CVPR'24	26.270	0.9196	1.965M	✔
UHDIP	arxiv'24	26.749	0.9281	0.81M	✔
WaveMamba	MM'24	27.350	0.9130	1.258M	✔
Ours	-	**27.729**	**0.9287**	1.215M	✔

Table 3: Quantitative results of low-light image enhancement.

Method	Venue	PSNR ↑	SSIM ↑	Param ↓	FS
FHDe^{2}Net	ECCV'20	20.338	0.7496	13.571M	✗
ESDNet	ECCV'22	22.119	0.7956	5.93M	✔
ESDNet-L	ECCV'22	22.422	0.7985	10.62M	✔
UHDformer	AAAI'24	21.968	0.8334	**0.3393M**	✔
UHDDIP	arxiv'24	22.068	0.8029	0.81M	✔
Ours	-	**23.242**	**0.8427**	1.456M	✔

Table 4: Quantitative results of image demoiréing.

Method	One Stage			Two Stage		
	Laz3	Bicubic	PS	AE	LMAR	Ours
PSNR	24.983	25.672	27.345	26.833	27.819	29.338
SSIM	0.747	0.7932	0.8409	0.8312	0.8341	0.8523

Table 5: Ablation on resampling methods.

Method	$\text{PSNR}_{\text{stage1}}$ ↑	$\text{PSNR}_{\text{stage2}}$ ↑	Param ↓	FLOPs ↓	FS
VAE	**35.313**	**30.014**	83.614M	445.307G	✗
VAE_{s}	31.234	25.613	5.243M	27.927G	✔
Ours	34.529	29.338	**1.061M**	**3.437G**	✔

Table 6: Ablation on FE-block. FLOPs are computed based on an input size of 256×256.

Method	w/o FE-block	w/o FAFE	w/o SFAD	w/o SFIM	Ours
PSNR	24.548	25.142	28.217	29.143	29.338
SSIM	0.758	0.763	0.843	0.849	0.8523

Table 7: Ablation on FE-block.

Method	w/o adapter	w/o DWT	w/o FAAM	Ours
PSNR	23.674	28.242	29.124	29.338
SSIM	0.742	0.8416	0.8473	0.8523

Table 8: Ablation on Wavelet Transform-based Adapter.

Comparison With Other Resamplers First, we compare our proposed FE-VAE with other resampling methods in Tab. 5. Here, Laz3 stands for Lanczos3, PS for pixel shuffle (Shi et al. 2016), and AE for autoencoder. One-stage refers to end-to-end training, while two-stage involves pre-

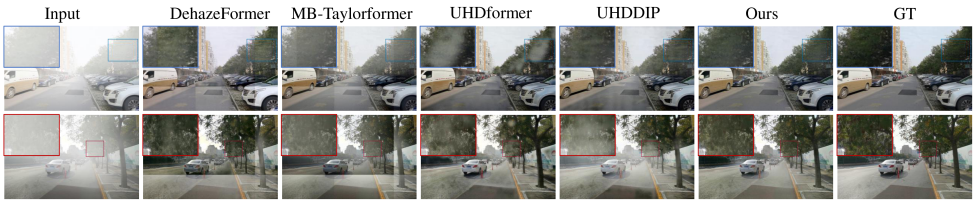

Figure 4: Image dehazing on UHD-Haze. FEVAE-UHD generates results with minimal haze residue.

Figure 5: Image deblurring on UHD-Blur. FEVAE-UHD generates results with the sharpest details.

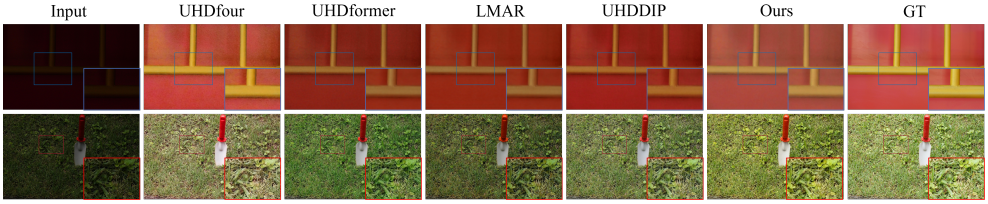

Figure 6: Low-light image enhancement on UHD-LL. FEVAE-UHD generates results with the most natural color fidelity.

training the resampler and freezing it during enhancer training. We ensure that all "resampler + enhancer" combinations have consistent parameters. The results show that FE-VAE significantly outperforms other resamplers, proving that its encoded latent space is both compact and informative, reducing computational costs while maintaining excellent performance.

Effect on FE-VAE We validate FE-VAE's efficiency and effectiveness by comparing it with the vanilla VAE and its lightweight version, VAEs, as shown in Tab. 6. While the vanilla VAE achieves the highest PSNR, it requires excessive parameters and FLOPs, making full-size inference impractical. The lightweight VAEs reduces parameters by simply lowering the number of channels and blocks, but suffers significant PSNR loss. In contrast, FE-VAE matches the vanilla VAE's performance while being more efficient, demonstrating the FE-block's effectiveness in enhancing parameter representation with frequency awareness.

Effect on FE-block We provide detailed ablation experiments of the FE-block in Tab. 7. The results demonstrate that introducing Frequency-aware Feature Extraction (FAFE) significantly enhances parameter representation capability. Additionally, the Space-frequency Adaptive Decomposition (SAFD) improves global channel awareness and filters features suited for FAFE processing. Finally, the fusion module effectively enhances interaction between the spatial and frequency branches, further boosting the performance.

Effect on Wavelet Transform-based Adapter Ablation experiments on the Wavelet Transform-based Adapter (WTA) are shown in Tab. 8. Without the adapter, the VAE struggles to bridge the domain gap, causing a significant performance drop. Replacing the wavelet transform with a convolutional layer and residual connection results in substantial high-frequency loss during encoding, leading to poor texture in the reconstruction. Additionally, the absence of the FAAM weakens spatial-frequency interaction between the WTA and FE-VAE branches, negatively affecting the final restoration.

Conclusion

We propose FEVAE-UHD, a universal UHD image restoration framework that combines frequency priors with VAE latent space representation for improved consistency and lower computational cost. To enable full-resolution inference on consumer-grade GPUs, we design a frequency-enhanced VAE with Fourier priors to boost parameter representation. Additionally, we introduce a wavelet-based adapter to restore high-frequency details and address domain gaps. Experiments on various UHD restoration tasks show that our approach outperforms state-of-the-art methods.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants 62225207, 62436008, 62422609 and 62276243.

References

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple Baselines for Image Restoration. *arXiv preprint arXiv:2204.04676*.

Chen, W.; Xu, H.; Li, Z.; Pei, D.; Chen, J.; Qiao, H.; Feng, Y.; and Wang, Z. 2019. Unsupervised anomaly detection for intricate kpis via adversarial training of vae. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, 1891–1899. IEEE.

Chen, X.; Li, Z.; Pu, Y.; Liu, Y.; Zhou, J.; Qiao, Y.; and Dong, C. 2023. A Comparative Study of Image Restoration Networks for General Backbone Network Design. *arXiv preprint arXiv:2310.11881*.

Deng, S.; Ren, W.; Yan, Y.; Wang, T.; Song, F.; and Cao, X. 2021. Multi-Scale Separable Network for Ultra-High-Definition Video Deblurring. In *ICCV*, 14010–14019.

Duan, Z.; Lu, M.; Ma, J.; Huang, Y.; Ma, Z.; and Zhu, F. 2024. QARV: Quantization-Aware ResNet VAE for Lossy Image Compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 436–450.

Ge, C.; Fu, X.; He, P.; Wang, K.; Cao, C.; and Zha, Z.-J. 2024. Neuromorphic Event Signal-Driven Network for Video De-raining. In *AAAI*, volume 38, 1878–1886.

Kingma, D. P. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023a. Efficient frequency domain-based transformers for high-quality image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5886–5895.

Kong, L.; Dong, J.; Ge, J.; Li, M.; and Pan, J. 2023b. Efficient Frequency Domain-Based Transformers for High-Quality Image Deblurring. In *CVPR*, 5886–5895.

Li, C.; Guo, C.-L.; Zhou, M.; Liang, Z.; Zhou, S.; Feng, R.; and Loy, C. C. 2023a. Embedding Fourier for Ultra-High-Definition Low-Light Image Enhancement. In *ICLR*.

Li, D.; Zhu, J.; Fu, X.; Guo, X.; Liu, Y.; Yang, G.; Liu, J.; and Zha, Z.-J. 2025. Noise-Assisted Prompt Learning for Image Forgery Detection and Localization. In *European Conference on Computer Vision*, 18–36. Springer.

Li, D.; Zhu, J.; Wang, M.; Liu, J.; Fu, X.; and Zha, Z.-J. 2023b. Edge-Aware Regional Message Passing Controller for Image Forgery Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.

Li, Y.; Fan, Y.; Xiang, X.; Demandolx, D.; Ranjan, R.; Timofte, R.; and Van Gool, L. 2023c. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18278–18289.

Peng, L.; Cao, Y.; Sun, Y.; and Wang, Y. 2024. Lightweight Adaptive Feature De-drifting for Compressed Image Classification. *IEEE Transactions on Multimedia*.

Qiu, Y.; Zhang, K.; Wang, C.; Luo, W.; Li, H.; and Jin, Z. 2023. MB-TaylorFormer: Multi-branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing. .

Recasens, A.; Kellnhofer, P.; Stent, S.; Matusik, W.; and Torralba, A. 2018. Learning to zoom: a saliency-based sampling layer for neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 51–66.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A. P.; Bishop, R.; Rueckert, D.; and Wang, Z. 2016. 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*, 1874–1883.

Sun, W.; and Chen, Z. 2020. Learned image downscaling for upscaling using content adaptive resampler. *IEEE Transactions on Image Processing*, 29: 4027–4040.

Wang, C.; Pan, J.; Wang, W.; Fu, G.; Liang, S.; Wang, M.; Wu, X.-M.; and Liu, J. 2024a. Correlation Matching Transformation Transformers for UHD Image Restoration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 5336–5344.

Wang, L.; Wang, C.; Pan, J.; Zhou, W.; Sun, X.; Wang, W.; and Su, Z. 2024b. Ultra-High-Definition Restoration: New Benchmarks and A Dual Interaction Prior-Driven Solution. arXiv:2406.13607.

Wang, L.; Wang, C.; Pan, J.; Zhou, W.; Sun, X.; Wang, W.; and Su, Z. 2024c. Ultra-High-Definition Restoration: New Benchmarks and A Dual Interaction Prior-Driven Solution. arXiv:2406.13607.

Xiao, J.; Feng, R.; Zhang, H.; Liu, Z.; Yang, Z.; Zhu, Y.; Fu, X.; Zhu, K.; Liu, Y.; and Zha, Z.-J. 2024. DreamClean: Restoring Clean Image Using Deep Diffusion Prior. In *The Twelfth International Conference on Learning Representations*.

Yu, R. 2020. A Tutorial on VAEs: From Bayes' Rule to Lossless Compression. arXiv:2006.10273.

Yu, W.; Huang, J.; Li, B.; Zheng, K.; Zhu, Q.; Zhou, M.; and Zhao, F. 2024. Empowering Resampling Operation for Ultra-High-Definition Image Enhancement with Model-Aware Guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25722–25731.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient Transformer for High-Resolution Image Restoration. In *CVPR*.

Zheng, Z.; and Jia, X. 2022. UHD Image Deblurring via Multi-scale Cubic-Mixer. arXiv:2206.03678.

Zheng, Z.; Ren, W.; Cao, X.; Hu, X.; Wang, T.; Song, F.; and Jia, X. 2021a. Ultra-High-Definition Image Dehazing via Multi-Guided Bilateral Learning. In *CVPR*, 16185–16194.

Zheng, Z.; Ren, W.; Cao, X.; Wang, T.; and Jia, X. 2021b. Ultra-high-definition image hdr reconstruction via collaborative bilateral learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4449–4458.

Zhou, L.; Cai, C.; Gao, Y.; Su, S.; and Wu, J. 2018. Variational autoencoder for low bit-rate image compression. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2617–2620.

Zhu, J.; Li, D.; Fu, X.; Yang, G.; Huang, J.; Liu, A.; and Zha, Z.-J. 2024a. Learning Discriminative Noise Guidance for Image Forgery Detection and Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7739–7747.

Zhu, Y.; Fu, X.; Zhang, Z.; Liu, A.; Xiong, Z.; and Zha, Z.-J. 2024b. Hue Guidance Network for Single Image Reflection Removal. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10): 13701–13712.

Zou, W.; Gao, H.; Yang, W.; and Liu, T. 2024. Wave-Mamba: Wavelet State Space Model for Ultra-High-Definition Low-Light Image Enhancement. arXiv:2408.01276.