Compression Artifacts Removal for VVC with Frequency Domain Mixture of Experts Network

Qijun Wang[1]*, **Kang Wang**[1], **Jun Wang**[1]

[1]Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Artificial Intelligence, Anhui University, Hefei, China
wangqijun308@163.com

Abstract

In recent years, lossy compression algorithms such as H.264/AVC, H.265/HEVC, and H.266/VVC have been proposed and widely applied in image and video encoding. However, these compression algorithms inevitably introduce various complex types of compression artifacts, which severely degrade image quality. Although existing methods have attempted to remove artifacts through filter design or probabilistic prior modeling, they are often effective only for specific types of artifacts, lacking generalization and adaptability. To address this, we propose a novel image compression artifacts removal model: ARMoE, which combines multiple frequency domain transformations with the Mixture of Experts (MoE). Considering the frequency distribution and energy distribution differences of images, we introduce various frequency domain transformations as expert branches and use the Sparse Activation Strategy to adaptively select the optimal frequency domain expert to suppress compression artifacts, achieving an efficient artifacts removal method. Furthermore, we reencode and decode multiple original uncompressed high-quality datasets, including DF2K and Kodak24, using the VTM-20.0 codec under the H.266/VVC standard, constructing a more challenging artifacts dataset. We conducted rigorous comparative experiments with current state-of-the-art image restoration methods and the results demonstrate that ARMoE exhibits outstanding image restoration capability.

Code — https://github.com/Kang341281X/ARMoE

1 Introduction

With the rapid development of communication technology, efficient communication compression technology is crucial for transmitting large amounts of image data. To address significant space overhead and limited bandwidth issues encountered during image data storage and transmission, lossy compression algorithms such as H.264/AVC (Sullivan and Wiegand 2005), H.265/HEVC (Sullivan et al. 2012), and H.266/VVC (Bross et al. 2021) have been proposed and widely adopted for image compression tasks. Lossy compression algorithms achieve efficient feature compression by quantizing redundant information in images. Although they help alleviate bandwidth constraints during data transmission, they inevitably introduce various compression artifacts, such as blocking effects, ringing phenomena, and blurring.

In recent years, deep learning-based network architectures have been progressively applied to various computer vision tasks, leading to the proliferation of numerous image restoration methods based on Convolutional Neural Networks (CNNs) (Zhang et al. 2017b, 2018b). Although CNNs have shown superior image restoration performance, their primary drawbacks lie in their limited receptive field and their predominant focus on modeling local patterns, making it challenging to effectively capture long-range dependencies and global contextual information. Although the Transformer was originally proposed in the field of Natural Language Processing (NLP), it has garnered widespread attention in the field of computer vision and has demonstrated excellent performance in various vision tasks (Chen et al. 2022; Liu et al. 2021). The Transformer can effectively capture long-range dependencies through self-attention, thereby achieving more precise contextual modeling and feature representation in image tasks.

Inspired by the superior performance of Transformers, many researchers have combined CNNs with Transformers to address the task of removing compression artifacts. Despite achieving significant progress, the following challenges persist: **1)** Multiple artifacts types generated during the compression process, such as blocking artifacts, ringing effects, and blurring, possess distinct spatial characteristics. Existing methods often perform well only on a specific type of artifacts, making it difficult to achieve synergistic and efficient removal of diverse artifacts types. **2)** A substantial portion of existing research primarily focuses on artifacts linked with the JPEG compression standard (Foi, Katkovnik, and Egiazarian 2007). There is a seriously inadequate in studies examining compression artifacts within the H.266/VVC standard, resulting in a scarcity of comprehensive and effective cross-standard artifacts removal frameworks and datasets. **3)** The computational complexity of Transformer-based models is $O(N^2 \cdot d)$ (where N represents the input sequence length and d is the channel dimension). Consequently, these models introduce substantial computational overhead when applied to high-resolution image inputs, posing a critical challenge to balancing computational

*Corresponding author: Qijun Wang.

efficiency with desired performance.

With the exponential growth in model and dataset scale, the Mixture of Experts (MoE) architecture (Jacobs et al. 1991; Yuksel, Wilson, and Gader 2012) has progressively emerged as a crucial architectural evolution for large language models and is now widely applied in foundation models (Lin et al. 2024; Dai et al. 2024). The core of MoE lies in its Sparse Activation Strategy, where each token activates only a subset of expert sub-networks. This approach significantly reduces computational overhead while maintaining powerful expressive capabilities, thereby achieving a more favorable balance between model performance and inference efficiency compared to traditional dense models. Furthermore, we observe that in the task of image compression artifacts removal, a single frequency domain transformation typically only addresses specific types of artifacts, struggling to cope with the complex and diverse characteristics of various artifacts. Consequently, inspired by the MoE architecture and frequency domain transformations, we propose treating multiple frequency domain transformations as distinct "experts". By leveraging MoE's Sparse Activation Strategy, these experts can be synergistically orchestrated to achieve fine-grained modeling and effective removal of complex types of compression artifacts.

While several public datasets for image compression artifacts removal exist, such as DIV2K (Timofte et al. 2017), Flickr2K (Lim et al. 2017), BSD500 (Arbelaez et al. 2010), and WED (Ma et al. 2016), these datasets are all based on the JPEG encoding and decoding method. There's still a significant lack of image datasets specifically addressing artifacts generated by the H.266/VVC standards. To fill this gap in current research, we reencode and decode five existing datasets: DIV2K, Flickr2K, Kodak24 (Franzen 1999), CBSD68 (Martin et al. 2001) and McMaster (Zhang et al. 2011), using the H.266/VVC encoding and decoding standard. This process created a more representative compression artifacts dataset named VVC-CAR.

Overall, our work makes three key contributions:

- We designed Artifacts Removal Mixture of Experts (ARMoE), specifically for removing image compression artifacts under the H.266/VVC codec standard. ARMoE integrates four distinct frequency domain transforms as expert modules. Using a Sparse Activation Strategy, the model adaptively activates the most suitable frequency domain expert, thereby enabling more targeted and effective artifacts removal operations.
- We construct the VVC-CAR dataset based on the H.266/VVC standard by reencode and decode five commonly used image datasets. Each subset includes uncompressed high-quality datasets and their compressed counterparts at four different quantization parameters levels.
- We systematically evaluated the ARMoE model for image compression artifacts removal and conducted a fairness study in lightweight image super-resolution. Experimental results show ARMoE achieved state-of-the-art performance in both tasks, demonstrating its superior image restoration capabilities.

2 Related Work

2.1 Filter-Based Methods

Early approaches focused on hand-crafted filter design to suppress compression artifacts. Zhai et al. (Zhai et al. 2008) combined DCT-domain regularization and spatial-domain adaptive filtering to enhance local correlations. Yoo et al. (Yoo, Choi, and Ra 2014) smoothed low-frequency DCT coefficients and applied group-based filtering to improve interblock consistency. However, these manually designed filters offer limited representational capacity, making it difficult to achieve satisfactory restoration quality.

2.2 Probabilistic Prior-Based Methods

Many traditional methods treat artifacts removal as an inverse problem, leveraging probabilistic priors for distortion correction. Sun et al. (Sun and Cham 2007) modeled compression-induced distortions as spatially correlated Gaussian noise and used a high-order Markov random field to represent clean images. Zhang et al. (Zhang et al. 2012) proposed a non-local artifacts removal method by exploiting inter-block similarity and image statistics in the transform domain. However, such prior-based methods often rely on assumptions about natural image statistics, which may not generalize well to complex or highly textured scenes, limiting their effectiveness.

2.3 Deep Learning-Based Methods

Deep learning-based methods aim to learn the mapping from compressed to original images, achieving remarkable progress in artifacts removal. ARCNN (Dong et al. 2015) as a pioneering work, employs a three-layer CNN to remove blocking and ringing artifacts. REDNet (Mao, Shen, and Yang 2016) introduces a deep encoder–decoder framework with multiple convolutional and deconvolutional layers for end-to-end restoration. DMCNN (Zhang et al. 2018a) leverages redundant information from both pixel and DCT domains to effectively eliminate banding artifacts.

Image compression artifacts are spatially diverse, leading to a complex array of artifact types where different regions may exhibit distinct distortion characteristics. Consequently, a single frequency domain transform often struggles to comprehensively address the intricate task of removing compression artifacts. To overcome this, we introduce multiple frequency domain transforms, each offering distinct advantages for specific artifacts types: Discrete Cosine Transform (DCT) excels at energy compaction, effectively managing block artifacts and boundary discontinuities. Fast Fourier Transform (FFT) is proficient in global frequency modeling, capturing interblock structural artifacts. Discrete Wavelet Transform (DWT) offers spatial-frequency localization, making it well-suited for removing edge blurring and recovering local structural details. Stationary Wavelet Transform (SWT) maintains spatial consistency by avoiding downsampling, making it ideal for restoring fine textures and preserving edge integrity.

Therefore, we use multiple frequency domain transformations and the Sparse Activation Strategy to eliminate complex types of compression artifacts in images.

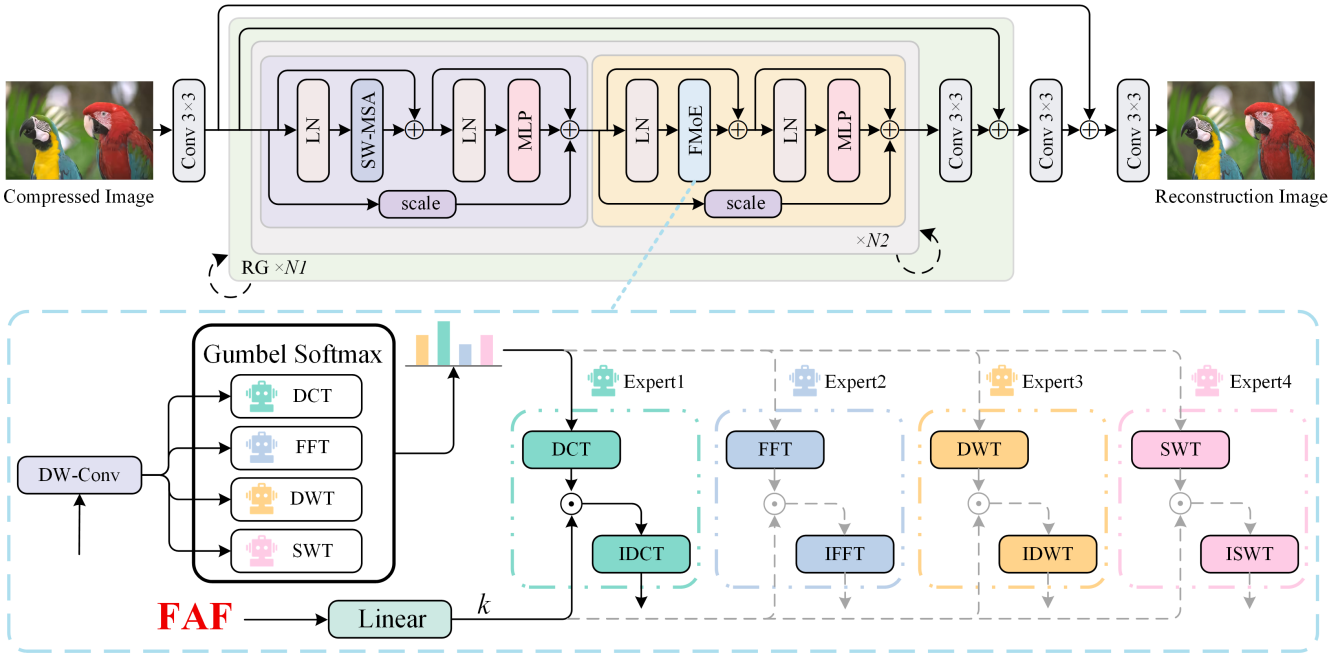

Figure 1: The network architecture of the Artifacts Removal Mixture of Experts (ARMoE) model is illustrated. In the Frequency Domain Mixture of Experts (FMoE) module, selected experts are indicated by solid black lines, while unselected experts are represented by dashed gray lines.

3 Methods

3.1 Architecture

The overall network of the proposed ARMoE consists of three main parts: shallow feature extraction, deep feature extraction, and image reconstruction, as illustrated in Fig. 1.

Initially, given a compressed low-quality (LQ) image as input $I_{LQ} \in \mathbb{R}^{H\times W\times 3}$, it's processed by a convolutional layer to generate the shallow features $F_S \in \mathbb{R}^{H\times W\times C}$. Notations H and W denote the height and width of the input image, while C represents the number of feature channels.

Subsequently, the shallow features F_S are fed into the deep feature extraction module to obtain deep features $F_D \in \mathbb{R}^{H\times W\times C}$. This module comprises $N1$ Residual Groups (RGs). Shifted Window Multi-head Self-Attention (SW-MSA) is effective not only in modeling long-range dependencies but also in perceiving structures across windows. Consequently, we integrate SW-MSA with the Frequency Domain Mixture of Experts (FMoE) to form the residual unit module of ARMoE. Each residual group incorporates $N2$ SW-MSA and FMoE blocks, processing data with an alternating network architecture. To further enhance feature representation capability, a convolutional layer is introduced at the end of each RG to refine the output features. In addition, a residual connection structure is employed within each RG to improve training stability.

Finally, the deep features are fed into the reconstruction module, where a convolutional operation maps the features to the output channels, generating the final reconstructed image $F_{Rec} \in \mathbb{R}^{H\times W\times 3}$.

3.2 Sparse Activation Strategy

The MoE (Jacobs et al. 1991; Yuksel, Wilson, and Gader 2012) is a deep learning architecture that processes input data using multiple expert networks and a gating mechanism. Its core lies in a Sparse Activation Strategy, which divides complex tasks into several subtasks. The most suitable expert subnetwork then handles each subtask, thereby significantly improving computational efficiency while ensuring model performance.

In an MoE, each expert network is an independent neural network with its own parameters. Let's denote $G(\cdot)$ as the output of the gating mechanism for input x, and $E_i(\cdot)$ as the output of the i-th expert for input x. The output Y of the MoE module can then be written as:

$$Y = \sum_{i=1}^{n} G(x)_i E_i(x), \tag{1}$$

Sparsity is one of the key properties of the gating network $G(\cdot)$. This design significantly reduces computational overhead and helps improve inference efficiency. The mathematical definition of the gating function is given as follows:

$$\begin{aligned} G(x) &= \text{Softmax}\left(\text{TopK}(H(x), k)\right), \\ \text{TopK}(v, k)_i &= \begin{cases} v_i, & \text{if } v_i \in \text{Top}_k(v), \\ -\infty, & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

The scoring function $H(\cdot)$ calculates the degree of matching between the input token and each expert, providing a weight vector of dimensions N. Each component of this vector represents the score for the corresponding expert. To achieve

a Sparse Activation Strategy, the $TopK(H(\cdot), k)$ operation is applied. This keeps only the top k experts with the highest scores, while assigning $-\infty$ to the remaining positions to block $Softmax$ activation. Subsequently, the Softmax function is applied to obtain the normalized gating vector $G(\cdot)$, which is then used to weight the outputs of the activated experts.

3.3 Frequency Domain Transform

Addressing the discontinuous spatial distribution of image compression artifacts, and drawing inspiration from vHeat (Wang et al. 2025) which employs 2D Discrete Fourier Transform (DFT) and its inverse (IDFT) to simulate heat conduction in the visual domain, we establish the transformation relationship between the spatial and frequency domains:

$$u(x, y, t) = \mathcal{F}^{-1}(\mathcal{F}(z_x, z_y) \cdot e^{-k\left(z_x^2 + z_y^2\right)t}), \qquad (3)$$

Here, $u(x, y, t)$ denotes the distribution intensity of image features at position (x, y) at time t. $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ represent the 2D Fourier Transform and 2D Inverse Fourier Transform, respectively. (z_x, z_y) indicates the frequency components of the image features at (x, y) in the frequency domain, and k is the thermal diffusivity.

Inspired by this, we view image compression as a process of spatial information degradation and artifact generation, which allows us to restore spatial coherence and effectively remove compression artifacts.

3.4 Frequency Domain Mixture of Experts

The overall architecture of FMoE closely resembles that of ViT. However, instead of applying the MoE to the feed forward neural network layers as is typical, we replace ViT's self-attention mechanism with our proposed FMoE, while keeping the rest of the framework intact. Existing research has already validated this framework's effectiveness in vision tasks. This alternative design not only boosts performance but also significantly reduces the computational overhead traditionally associated with attention mechanisms.

In this paper, we jointly train four pairs of frequency domain transforms as independent expert networks. These include: DCT-IDCT (Discrete Cosine Transform and its Inverse), FFT-IFFT (Fast Fourier Transform and its Inverse), DWT-IDWT (Discrete Wavelet Transform and its Inverse), and SWT-ISWT (Stationary Wavelet Transform and its Inverse). As illustrated in Fig. 1, we first use Depthwise Separable Convolutions (DW-Conv) to expand the image's spatial features along the channel dimension, resulting in a multi-channel feature representation denoted as U_0. To dynamically select frequency domain transform experts, we introduce a Gumbel Softmax-based strategy network. By incorporating Gumbel noise, we approximate the discrete selection process as a continuous and differentiable operation. This enables a differentiable expert selection mechanism, allowing for end-to-end training of the entire model. Inspired by position embeddings in ViT, we also introduce and randomly initialize Frequency-Aware Factors (FAF), which serve as a baseline decay map with the same shape as the input features. Additionally, FAF generates a learnable frequency decay exponent k via a $Linear$ layer, used to dynamically adjust FAF's decay characteristics. Specifically, by raising the FAF to the power determined by k, we generate the final FAF that accurately reflects the degree of artifact interference experienced by different frequency components of image features after compression. This factor is subsequently multiplied by the transformed frequency domain feature representation to yield the frequency domain modulated result. Through the corresponding inverse frequency domain transform, the result is converted back to the spatial domain, thereby achieving the goal of regulating frequency domain decay and efficiently suppressing compression artifacts across different frequency bands.

The diverse and complex nature of compression artifacts suggests that a single DCT-IDCT transform might not be optimal for effective modeling. Therefore, we explore the use of multiple frequency domain transforms to perform adaptive frequency operations and artifact suppression. This approach allows us to account for varying artifact types and regional frequency differences.

In physics, when partial differential equations describe phenomena within a finite region, boundary conditions must be explicitly defined for that region to ensure a unique and physically reasonable solution. Similarly, in the visual domain, image data is inherently spatially constrained, and semantic information doesn't propagate infinitely beyond image blocks. This naturally creates implicit boundaries. The presence of these "implicit boundaries" makes frequency domain modeling of features within an image block both logical and physically interpretable. Therefore, the output feature U_t, ultimately mapped back to the spatial domain, can be expressed as:

$$U_t = \mathcal{T}^{-1}\left(\mathcal{T}(U_0)\, e^{-k\left(z_x^2 + z_y^2\right)t}\right). \qquad (4)$$

Here, $\mathcal{T}$ can be any one of the frequency domain transform operations: DCT, FFT, DWT or SWT, while $\mathcal{T}^{-1}$ denotes its inverse transform operation.

Building on the theoretical foundation of frequency transforms, FMoE can adaptively select the most suitable frequency expert for processing based on the content characteristics of the input region. In low-frequency regions with sparse backgrounds, compression typically introduces minimal noticeable artifacts, and images often exhibit large, uniform areas with consistent properties. When such an image is divided into blocks, artifacts are frequently confined to individual local blocks. In these scenarios, the network's focus can be limited to a single block, allowing for effective feature restoration by applying local frequency transforms such as DCT, DWT or SWT solely within that block. However, in high-frequency regions with dense structures, compression more readily generates noticeable artifacts like block effects, ringing artifacts or contour distortions, which typically span multiple blocks. Therefore, considering the contextual relationships between blocks is crucial for improving restoration. FFT transforms can perform global frequency domain modeling of the image, enabling cross-block structural awareness and globally consistent artifacts removal.

Baseline	DCT	FFT	DWT	SWT	Kodak24		McMaster		CBSD68	
					PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✓					32.75	0.8630	33.59	0.8976	31.63	0.8585
✓	✓				32.77	0.8632	33.63	0.8976	31.63	0.8585
✓	✓	✓			32.83	0.8639	33.76	0.8987	31.66	0.8592
✓	✓	✓	✓		32.84	0.8641	33.76	0.8987	31.67	0.8593
✓	✓	✓	✓	✓	**32.87**	**0.8644**	**33.81**	**0.8995**	**31.69**	**0.8597**

Table 1: Ablation study of frequency domain transform. Best results are shown in bold. Quantization parameter (QP) = 37.

Charbonnier	L1	Kodak24		McMaster	
		PSNR	SSIM	PSNR	SSIM
✓		32.83	0.8639	33.77	0.8985
	✓	**32.87**	**0.8644**	**33.81**	**0.8995**

Table 2: Ablation study of loss functions. Best results are shown in bold. Quantization parameter (QP) = 37.

Model	Kodak24		McMaster		CBSD68	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
FAF w/o	32.67	0.8607	33.51	0.8949	31.53	0.8562
FAF w	**32.87**	**0.8644**	**33.81**	**0.8995**	**31.69**	**0.8597**

Table 3: Ablation study of FAF. Best results are shown in bold. Quantization parameter (QP) = 37.

4 Experiments

4.1 Experimental Settings

For the image compression artifacts removal task, we utilize VVC-CAR as our dataset. Specifically, we processed the original uncompressed high-quality datasets using the FFmpeg decoder. The process involved converting the datasets from RGB to YUV420P format. Subsequently, we reencode and decode them using the latest VTM-20.0 codec under the H.266/VVC standard, employing an All-Intra (AI) configuration with four different quantization parameters (QP=22, 27, 32, 37). Finally, the data was converted back to RGB format. Within the VVC-CAR dataset, we utilized the DF2K dataset for model training, and Kodak24, McMaster, and CBSD68 for model testing. BD-rate reduction were calculated using the 16 common test sequences (Class B-E) recommended by JVET (Zhang et al. 2019). We employed a consistent training strategy, training ARMoE alongside current mainstream image restoration methods. For each method, we obtained four models corresponding to different quantization levels. The batch size was set to 2, and the channel dimension was set to 180.

For the lightweight image super-resolution task, we used DIV2K (comprising 900 images) as the training dataset, and Set5 (Bevilacqua et al. 2012) and BSD100 (Martin et al. 2001) as the test datasets. Low-resolution images were generated from ground-truth images by "bicubic" downsampling in MATLAB. We trained a lightweight version of ARMoE for a 2 scaling factor. To ensure fairness in our experiments, given that the comparison methods in the image compression artifact removal task were re-trained by us, we also re-trained these methods on the common lightweight image super-resolution task. For this task, we adjusted the batch size to 16 and set the channel dimension to 48.

The following general configurations were adopted during the training process. Data augmentation was performed by applying horizontal flips and random rotations of 90°, 180°, and 270°. Additionally, original images were cropped into 64×64 patches for training. Adam (Diederik 2014) was employed as the optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The initial learning rate was set to 2×10^{-4} for 500K training iterations, with the learning rate being halved at specific training milestones. $L1$ loss was used as the loss function, and the number of attention heads was set to 6. Our model was trained and tested on a single NVIDIA RTX 3090 GPU. Test results were evaluated using two metrics: Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) (Wang et al. 2004). In this work, PSNR and SSIM were calculated on the Y channel of the YCbCr space and reported as quantitative evaluation metrics.

4.2 Ablation Study

Frequency Domain Transform. To validate the effectiveness of different frequency domain transform combinations within the FMoE module, we conducted systematic ablation experiments, with results presented in Table 1. The "baseline" denotes the network configuration where the FMoE architecture is removed, retaining only the SW-MSA module. Experimental results indicate a consistent improvement in both PSNR and SSIM with the increasing inclusion of frequency domain transforms. Ultimately, optimal performance across all three test datasets was achieved when all four frequency domain transform experts were incorporated. These findings demonstrate that different frequency domain transform experts offer complementary advantages in handling various artifact types, and their joint modeling enables more precise frequency regulation and artifact removal.

Loss Function. In previous JPEG compression artifact removal tasks, the $Charbonnier$ loss function was widely adopted due to its smoothness. However, its excessive smoothness often leads to lower sensitivity in reconstructing fine image details, potentially resulting in insufficient detail recovery. As shown in Table 2, employing the $L1$ loss function yielded better performance in the Kodak24 and McMaster datasets. This indicates that $L1$ loss offers a stronger advantage in preserving and reconstructing image details.

Frequency-Aware Factors (FAF). FAF represent the degree of artifact interference experienced by different fre-

QP	Method	Publish	Param	Kodak24		McMaster		CBSD68	
				PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
22	SwinIR	ICCVW 2021	16.3M	42.97	0.9786	43.46	0.9823	43.15	0.9857
	Restormer	CVPR 2022	26.1M	42.92	0.9784	43.40	0.9821	43.12	0.9856
	DAT	ICCV 2023	14.5M	42.98	0.9787	43.47	0.9823	43.16	0.9857
	HAT	CVPR 2023	20.5M	42.98	0.9787	43.48	0.9823	43.16	0.9857
	RGT	ICLR 2024	9.9M	42.99	0.9787	43.50	0.9824	43.17	0.9858
	DRCT	CVPR 2024	13.8M	42.95	0.9786	43.44	0.9820	43.15	0.9857
	MambaIR	ECCV 2024	20.3M	42.96	0.9786	43.45	0.9823	43.14	0.9857
	ARMoE	-	25.6M	**43.06**	**0.9789**	**43.54**	**0.9825**	**43.20**	**0.9859**
27	SwinIR	ICCVW 2021	16.3M	39.47	0.9597	40.13	0.9675	38.92	0.9654
	Restormer	CVPR 2022	26.1M	39.39	0.9593	40.04	0.9671	38.87	0.9651
	DAT	ICCV 2023	14.5M	39.47	0.9598	40.13	0.9675	38.91	0.9654
	HAT	CVPR 2023	20.5M	39.46	0.9598	40.11	0.9675	38.91	0.9654
	RGT	ICLR 2024	9.9M	39.47	0.9599	40.13	0.9676	38.91	0.9655
	DRCT	CVPR 2024	13.8M	39.40	0.9594	40.05	0.9671	38.87	0.9652
	MambaIR	ECCV 2024	20.3M	39.43	0.9596	40.09	0.9673	38.89	0.9653
	ARMoE	-	25.6M	**39.56**	**0.9603**	**40.23**	**0.9680**	**38.97**	**0.9657**
32	SwinIR	ICCVW 2021	16.3M	36.01	0.9234	36.87	0.9410	35.10	0.9267
	Restormer	CVPR 2022	26.1M	35.96	0.9230	36.83	0.9409	35.08	0.9265
	DAT	ICCV 2023	14.5M	36.03	0.9239	36.91	0.9416	35.12	0.9270
	HAT	CVPR 2023	20.5M	36.03	0.9238	36.89	0.9414	35.11	0.9270
	RGT	ICLR 2024	9.9M	36.05	0.9241	36.93	0.9417	35.13	0.9271
	DRCT	CVPR 2024	13.8M	35.98	0.9231	36.84	0.9408	35.08	0.9266
	MambaIR	ECCV 2024	20.3M	36.01	0.9236	36.88	0.9414	35.11	0.9271
	ARMoE	-	25.6M	**36.11**	**0.9246**	**36.99**	**0.9423**	**35.16**	**0.9276**
37	SwinIR	ICCVW 2021	16.3M	32.74	0.8619	33.66	0.8967	31.61	0.8575
	Restormer	CVPR 2022	26.1M	32.75	0.8625	33.69	0.8976	31.63	0.8584
	DAT	ICCV 2023	14.5M	32.82	0.8635	33.77	0.8987	31.67	0.8590
	HAT	CVPR 2023	20.5M	32.83	0.8636	33.77	0.8986	31.66	0.8589
	RGT	ICLR 2024	9.9M	32.84	0.8641	33.78	0.8990	31.67	0.8594
	DRCT	CVPR 2024	13.8M	32.78	0.8627	33.70	0.8975	31.64	0.8582
	MambaIR	ECCV 2024	20.3M	32.78	0.8633	33.71	0.8982	31.65	0.8590
	ARMoE	-	25.6M	**32.87**	**0.8644**	**33.81**	**0.8995**	**31.69**	**0.8597**

Table 4: Quantitative comparison with state-of-the-art methods. Best results are shown in bold.

Figure 2: Visual comparison results for challenging cases with a Quantization parameter (QP) = 37.

quency components of image features after compression, and they are a critical factor influencing the effectiveness of frequency domain modeling. As presented in Table 3, introducing FAF into our frequency domain modeling consistently improved performance metrics across all test datasets. This indicates that FAF effectively suppresses compression artifact interference, thereby enhancing the modeling capability of various frequency components.

4.3 Comparison with State-of-the-Art Methods

Quantitative Results. To comprehensively validate ARMoE's effectiveness in image compression artifacts removal, we re-trained and tested it against current state-of-the-art image restoration methods (including SwinIR (Dosovitskiy et al. 2020), Restormer (Zamir et al. 2022), DAT (Chen et al. 2023b), HAT (Chen et al. 2023a), RGT (Chen et al. 2024), DRCTP (Hsu, Lee, and Chou 2024), and MambaIR (Guo et al. 2024)) on the VVC-CAR dataset. We also included the number of parameters (Param) as an efficiency metric. Table 4 presents the results for the image compression artifact removal task across four quantization levels. Our ARMoE consistently outperforms all compared methods on all datasets and at all four quantization parameters. Specifically, compared to the ViT-based SwinIR and Mamba-based

QP	Methods	Class B Y(%)	Class C Y(%)	Class D Y(%)	Class E Y(%)
22	SwinIR	-29.18	-25.39	-27.74	-32.29
	Restormer	-28.57	-24.76	-27.46	-31.77
	DAT	-28.77	-25.50	-28.13	-32.74
	HAT	-28.69	-25.51	-28.23	-32.75
	RGT	-28.80	-25.76	-28.64	-32.00
	DRCT	-28.93	-24.88	-27.19	-32.51
	MambaIR	-29.17	-25.32	-28.30	-32.80
	ARMoE	**-29.48**	**-26.58**	**-29.83**	**-33.98**
27	SwinIR	-19.63	-15.15	-16.31	-24.26
	Restormer	-18.86	-14.76	-16.25	-22.68
	DAT	-19.84	-15.70	-17.21	-24.23
	HAT	-19.58	-15.60	-16.95	-24.45
	RGT	-16.87	-15.57	-16.91	-22.65
	DRCT	-18.78	-14.34	-15.24	-23.56
	MambaIR	-19.05	-15.02	-17.02	-23.25
	ARMoE	**-21.13**	**-17.47**	**-19.15**	**-26.89**
32	SwinIR	-15.43	-10.02	-10.81	-19.25
	Restormer	-13.75	-9.53	-10.94	-17.39
	DAT	-14.33	-10.68	-11.83	-19.31
	HAT	-15.03	-10.55	-11.42	-19.11
	RGT	-14.58	-11.03	-12.10	-18.76
	DRCT	-14.49	-9.29	-10.03	-18.17
	MambaIR	-14.31	-10.40	-11.88	-18.67
	ARMoE	**-17.06**	**-12.36**	**-13.31**	**-22.01**
37	SwinIR	-9.53	-5.60	-6.68	-12.56
	Restormer	-9.45	-6.31	-8.51	-12.15
	DAT	-9.81	-7.72	-9.28	-15.20
	HAT	-11.47	-7.85	-9.24	-15.81
	RGT	-10.34	-8.22	-9.74	-14.30
	DRCT	-10.47	-6.58	-7.83	-13.96
	MambaIR	-9.43	-6.91	-9.25	-14.07
	ARMoE	**-12.55**	**-9.02**	**-10.25**	**-17.23**

Table 5: Quantitative comparison on BD-rate reduction with state-of-the-art methods. Best results are shown in bold.

MambaIR, our ARMoE achieves significant gains on the McMaster dataset (QP=37), yielding performance improvements of 0.13 dB and 0.09 dB, respectively. Collectively, these quantitative results demonstrate that comprehensively utilizing multiple types of frequency domain transforms can effectively enhance image reconstruction quality.

Qualitative Analysis. We present visual comparison results for several challenging scenarios in Fig. 2. Previous methods often suffer from pervasive blurring, distortion, or inaccurate texture recovery. In contrast, our method effectively mitigates artifacts, preserving more structural integrity and finer details. This superior performance is primarily attributed to our approach's enhanced representational power, achieved by adaptively selecting the optimal frequency domain transform for image detail reconstruction via the Sparse Activation Strategy.

4.4 Comparison on BD-Rate Reduction

To comprehensively evaluate the BD-rate reduction of ARMoE against other methods across varying compression strengths, we tested it on 16 classic video test sequences (Class B–E). We used the Y channel, which contains luminance information, as the evaluation metric, and the experimental results are presented in Table 5. As the quantization level progressively decreases, all methods show improved bitrate savings, with ARMoE consistently achieving the best results across all quantization levels. This demonstrates ARMoE's ability to fully leverage its multiple frequency domain modeling advantages, adapt to complex compression artifact types, and achieve stable, efficient compression artifact removal and visual quality restoration in both high-compression (QP=37) and low-compression (QP=22) scenarios.

Method	Param	Set5 PSNR	Set5 SSIM	BSD100 PSNR	BSD100 SSIM
CARN	1,592K	37.76	0.9590	32.09	0.8978
SwinIR	910K	38.14	0.9611	32.31	0.9012
DIIN	726K	38.06	0.9610	32.20	0.8998
MambaIR	905K	38.13	0.9610	32.31	0.9013
ARMoE	845K	**38.23**	**0.9615**	**32.35**	**0.9019**

Table 6: Quantitative comparison of Lightweight Image Super-Resolution with a scaling factor of 2×.

4.5 Comparison on Lightweight Image Super-Resolution

To evaluate the effectiveness of ARMoE in lightweight image super-resolution tasks, we compare ARMoE with CARN (Ahn, Kang, and Sohn 2018), DIIN (Jin et al. 2024), SwinIR, and MambaIR, using a scaling factor of 2. The results in Table 6 show that our ARMoE outperforms the state-of-the-art method MambaIR while using fewer parameters. For example, our ARMoE outperforms MambaIR by 0.1 dB on the Kodak24 dataset with 60K fewer parameters. This experiment verifies the fairness and effectiveness of the proposed method.

5 Conclusion

In this paper, we propose Artifacts Removal Mixture of Experts (ARMoE), a novel model that effectively addresses the complex problem of diverse image compression artifacts by integrating four frequency domain transforms with a Mixture of Experts (MoE). Specifically, ARMoE transforms image features into the frequency domain and adaptively selects the optimal frequency domain expert for modeling using MoE's Sparse Activation Strategy. This selection is guided by factors such as the frequency distribution characteristics of different image regions and whether image blocks cross window boundaries, enabling precise artifact removal. Furthermore, to create a more challenging scenario for compression artifacts tasks, we developed the VVC-CAR dataset. This dataset was generated by reencode and decode several original uncompressed high-quality datasets, including DF2K and Kodak24, using the VTM-20.0 codec under the H.266/VVC standard. Extensive experiments indicate that ARMoE outperforms previous methods, offering a novel solution for image compression artifacts removal.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62472001; in part by the Natural Science Foundation of Anhui Province under Grant 2408085MF174 and Grant 2108085MF193; in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2022-034; and in part by the Key Program of Natural Science Project of Educational Commission of Anhui Province under Grant 2022AH050088.

References

Ahn, N.; Kang, B.; and Sohn, K.-A. 2018. Fast, Accurate, and Lightweight Super-Resolution with Cascading Residual Network. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Arbelaez, P.; Maire, M.; Fowlkes, C.; and Malik, J. 2010. Contour Detection and Hierarchical Image Segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 33(5): 898–916.

Bevilacqua, M.; Roumy, A.; Guillemot, C. M.; and Alberi-Morel, M.-L. 2012. Low-Complexity Single-Image Super-Resolution Based on Nonnegative Neighbor Embedding. In *British Machine Vision Conference*.

Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G. J.; and Ohm, J.-R. 2021. Overview of the Versatile Video Coding (VVC) Standard and Its Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10): 3736–3764.

Chen, Q.; Wu, Q.; Wang, J.; Hu, Q.; Hu, T.; Ding, E.; Cheng, J.; and Wang, J. 2022. Mixformer: Mixing Features Across Windows and Dimensions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5249–5259.

Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; and Dong, C. 2023a. Activating More Pixels in Image Super-Resolution Transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22367–22377.

Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; and Yang, X. 2024. Recursive Generalization Transformer for Image Super-Resolution. In *ICLR*.

Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; and Yu, F. 2023b. Dual Aggregation Transformer for Image Super-Resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12312–12321.

Dai, D.; Deng, C.; Zhao, C.; Xu, R.; Gao, H.; Chen, D.; Li, J.; Zeng, W.; Yu, X.; Wu, Y.; et al. 2024. Deepseekmoe: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. *arXiv preprint arXiv:2401.06066*.

Diederik, K. 2014. Adam: A Method for Stochastic Optimization. *(No Title)*.

Dong, C.; Deng, Y.; Loy, C. C.; and Tang, X. 2015. Compression Artifacts Reduction by a Deep Convolutional Network. In *Proceedings of the IEEE international conference on computer vision*, 576–584.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.

Foi, A.; Katkovnik, V.; and Egiazarian, K. 2007. Pointwise Shape-Adaptive DCT for High-Quality Denoising and Deblocking of Grayscale and Color Images. *IEEE transactions on image processing*, 16(5): 1395–1411.

Franzen, R. 1999. Kodak Lossless True Color Image Suite. *source: http://r0k. us/graphics/kodak*.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2024. Mambair: A Simple Baseline for Image Restoration With State-Space Model. In *European conference on computer vision*, 222–241. Springer.

Hsu, C.-C.; Lee, C.-M.; and Chou, Y.-S. 2024. DRCT: Saving Image Super-Resolution Away from Information Bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 6133–6142.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive Mixtures of Local Experts. *Neural computation*, 3(1): 79–87.

Jin, H.; Gao, G.; Li, J.; Guo, Z.; and Yu, Y. 2024. Efficient Dual-Branch Information Interaction Network for Lightweight Image Super-Resolution. *IEEE Transactions on Instrumentation and Measurement*, 73: 1–11.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1833–1844.

Lim, B.; Son, S.; Kim, H.; Nah, S.; and Mu Lee, K. 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 136–144.

Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Huang, J.; Zhang, J.; Pang, Y.; Ning, M.; et al. 2024. Moe-Llava: Mixture of Experts for Large Vision-Language Models. *arXiv preprint arXiv:2401.15947*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ma, K.; Duanmu, Z.; Wu, Q.; Wang, Z.; Yong, H.; Li, H.; and Zhang, L. 2016. Waterloo Exploration Database: New Challenges for Image Quality Assessment Models. *IEEE Transactions on Image Processing*, 26(2): 1004–1016.

Mao, X.; Shen, C.; and Yang, Y.-B. 2016. Image Restoration Using Very Deep Convolutional Encoder-Decoder Networks With Symmetric Skip Connections. *Advances in neural information processing systems*, 29.

Martin, D.; Fowlkes, C.; Tal, D.; and Malik, J. 2001. A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics. In *Proceedings eighth IEEE in-*

ternational conference on computer vision. ICCV 2001, volume 2, 416–423. IEEE.

Sullivan, G. J.; Ohm, J.-R.; Han, W.-J.; and Wiegand, T. 2012. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transactions on circuits and systems for video technology*, 22(12): 1649–1668.

Sullivan, G. J.; and Wiegand, T. 2005. Video Compression-From Concepts to the H. 264/AVC Standard. *Proceedings of the IEEE*, 93(1): 18–31.

Sun, D.; and Cham, W.-K. 2007. Postprocessing of Low Bit-Rate Block DCT Coded Images Based on a Fields of Experts Prior. *IEEE Transactions on Image Processing*, 16(11): 2743–2751.

Timofte, R.; Agustsson, E.; Van Gool, L.; Yang, M.-H.; and Zhang, L. 2017. Ntire 2017 Challenge on Single Image Super-Resolution: Methods and Results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 114–125.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. *Advances in neural information processing systems*, 30.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wang, Z.; Liu, Y.; Tian, Y.; Liu, Y.; Wang, Y.; and Ye, Q. 2025. Building Vision Models Upon Heat Conduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 9707–9717.

Yoo, S. B.; Choi, K.; and Ra, J. B. 2014. Post-Processing for Blocking Artifact Reduction Based on Inter-Block Correlation. *IEEE Transactions on Multimedia*, 16(6): 1536–1548.

Yuksel, S. E.; Wilson, J. N.; and Gader, P. D. 2012. Twenty Years of Mixture of Experts. *IEEE transactions on neural networks and learning systems*, 23(8): 1177–1193.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhai, G.; Zhang, W.; Yang, X.; Lin, W.; and Xu, Y. 2008. Efficient Deblocking With Coefficient Regularization, Shape-Adaptive Filtering, and Quantization Constraint. *IEEE Transactions on Multimedia*, 10(5): 735–745.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017a. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE transactions on image processing*, 26(7): 3142–3155.

Zhang, K.; Zuo, W.; Gu, S.; and Zhang, L. 2017b. Learning Deep CNN Denoiser Prior for Image Restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3929–3938.

Zhang, L.; Wu, X.; Buades, A.; and Li, X. 2011. Color Demosaicking by Local Directional Interpolation and Nonlocal Adaptive Thresholding. *Journal of Electronic imaging*, 20(2): 023016–023016.

Zhang, X.; Xiong, R.; Ma, S.; and Gao, W. 2012. Reducing Blocking Artifacts in Compressed Images via Transform-Domain Non-Local Coefficients Estimation. In *2012 IEEE International Conference on Multimedia and Expo*, 836–841. IEEE.

Zhang, X.; Yang, W.; Hu, Y.; and Liu, J. 2018a. DMCNN: Dual-Domain Multi-Scale Convolutional Neural Network for Compression Artifacts Removal. In *2018 25th IEEE international conference on image processing (icip)*, 390–394. IEEE.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018b. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In *Proceedings of the European conference on computer vision (ECCV)*, 286–301.

Zhang, Y.; Li, K.; Li, K.; Zhong, B.; and Fu, Y. 2019. Residual Non-Local Attention Networks for Image Restoration. *arXiv preprint arXiv:1903.10082*.