# Hierarchical Frequency-Guided Alignment Transformer for Compressed Video Quality Enhancement

**Liuhan Peng[1], Shuai Li[1]*, Yanbo Gao[1], Mao Ye[2], Chong Lv[1]**

[1]Shandong University
[2]University of Electronic Science and Technology of China
pengliuhan@gmail.com, {shuaili, ybgao}@sdu.edu.cn, cvlab.uestc@gmail.com, chonglv@mail.sdu.edu.cn

## Abstract

During the video encoding process, the original spatial domain signal is first transformed into the frequency domain, followed by quantization and compression. As a result, the quality degradation in compressed videos primarily stems from distortions in the frequency domain information. However, existing video enhancement methods typically directly fuse information from adjacent frames in the spatial domain, making it difficult for models to effectively compensate for frequency domain distortions, which leads to suboptimal detail restoration. To address this issue, we propose a Hierarchical Frequency-Guided Alignment Transformer. Additionally, by analyzing the characteristics of the frequency domain, we find that different frequency bands exhibit both correlations and a certain degree of independence. Based on this, we introduce a Frequency-Aware Transformer module that employs a combination of independent and mixed processing to optimize information exchange across different frequency domains, effectively mitigating cross-interference from irrelevant information. Experimental results demonstrate that, compared to existing methods, our approach achieves state-of-the-art performance in objective metrics (PSNR/SSIM), perceptual quality (LPIPS), and subjective visual effects, while reducing model complexity.

**Code** — https://github.com/pengliuhan/HFGAT

## Introduction

Over the past decade, the rapid growth of online video platforms and streaming services has significantly increased global video consumption. According to the Cisco Visual Networking Index (Forecast et al. 2019), video content now accounts for over 75% of global internet traffic, with 4K/8K ultra-high-definition formats becoming popular. Although modern video coding standards such as High Efficiency Video Coding (HEVC) (Sullivan et al. 2012) and Versatile Video Coding (VVC) (Bross et al. 2021) have improved compression efficiency, the trade-off between bitrate reduction and visual quality remains a challenge. The video compression process takes a hybrid coding framework, including prediction, transform coding, quantization, entropy coding, etc. Among them, the transform coding applies DCT or DST transforms to convert spatial blocks into frequency coefficients, and then the quantization discretizes the frequency components with the quantization step. This quantization step irreversibly discards information at different frequencies, especially the high-frequency details, introducing blocking artifacts. These distortions not only diminish perceptual quality but also adversely impact subsequent computer vision tasks. Therefore, using video enhancement techniques to address information losses has become important.

Existing video enhancement methods are predominantly categorized into two primary groups: single-frame-based and multi-frame-based methods. Single-frame-based methods (Yue et al. 2022) take the current frame as input and focus solely on spatial feature enhancement, while neglecting the temporal information inherent in adjacent frames. They are suitable for all intra coding configurations and for the low-delay or random-access coding configurations. Such a setting, only using the spatial information of the current frame, limits the model's performance. In contrast, multi-frame-based methods (Zhu et al. 2024a; Liu, Zhou, and Xiao 2022; Ding et al. 2021; Xu et al. 2021; Zhang et al. 2025a; Dong et al. 2024) integrate spatiotemporal information from adjacent frames, enabling better exploitation of temporal correlations in videos and thus achieving superior enhancement results. To leverage information from adjacent frames, early multi-frame enhancement methods primarily employed multiscale attentions (Luo et al. 2021) for inter-frame information fusion. However, due to the presence of complex motions, these methods face challenges in accurately aligning and effectively utilizing information from adjacent frames. Due to the ability of deformable convolution to adaptively adjust the sampling position of convolution kernels, researchers have started using deformable convolution (Zhang et al. 2022; He et al. 2025; Deng et al. 2024; Zhang et al. 2025b, 2024; Wang et al. 2025) for multi-frame alignment, significantly improving feature alignment capability. To enhance global modeling capabilities, Transformer (Yu et al. 2024) has also been introduced to video enhancement tasks, to capture the long-range spatiotemporal dependencies through self-attention mechanisms.

However, most of the existing multi-frame-based methods only fuse information from adjacent frames in the spatial domain and do not consider leveraging neighboring frame information in the frequency domain. Especially for the com-

*Corresponding author.

pressed video enhancement, the video reconstruction loss primarily results from the loss of frequency domain information due to the transform coding followed by quantization. Without exploring the frequency domain information, these approaches struggle to effectively capture and compensate for the frequency domain distortions, which may lead to insufficient restoration of video details after enhancement.

Moreover, we conducted an in-depth analysis of frequency domain features based on the wavelet transform and identified two key characteristics: (1) The data distribution of the different sub-bands is significantly different, including the data range and the distribution shape, as illustrated in Figs. 1(a) and (b). Therefore, directly mixing and inputting them into the network may lead to an overemphasis on certain frequency information or an insufficient exploration of others. (2) On the other hand, as shown in Fig. 2, the different sub-bands after wavelet transform exhibit a certain degree of independence (yellow boxes) while also demonstrating interdependence (red boxes). Therefore, in addition to focusing on the correlations between sub-bands and leveraging this information effectively, it is also necessary to conduct an in-depth analysis of the local structures within each sub-band to minimize interference among sub-bands.

To address the aforementioned issues and enhance the compressed video with distortion created in the transform domain, we propose a Hierarchical Frequency-Guided Alignment Transformer network. A Frequency-Aware Transformer module is developed, with a Frequency-Independent Transformer block and a Frequency-Mixed Transformer block. This design enables both independent processing within individual frequency sub-bands and inter-band information exchange, thereby strengthening the integration of spectral information while mitigating interference from irrelevant frequency components. Given the significant distribution differences among sub-bands, batch normalization is used after frequency domain decomposition instead of layer normalization to preserve the structural characteristics of different sub-bands while alleviating training difficulties caused by distribution disparities. Additionally, our architecture features a hierarchical structure with skip connections, thereby exploring long-range dependencies and multi-scale information. The main contributions of this paper are summarized as follows:

- We are the first to reveal two key characteristics of wavelet frequency sub-bands, and these findings provide a theoretical foundation for subsequent network design.
- Based on these characteristics, we carefully designed the Frequency-Aware Transformer module.
- The proposed method was comprehensively evaluated from three perspectives: subjective assessment, objective metrics, and perceptual quality. The results demonstrate that the method has achieved state-of-the-art performance.

## Related Work

### Single-Frame-based Video Enhancement

Early methods for video enhancement primarily focused on single-frame processing and achieved a series of progressive results. Among these, a representative example is the shallow AR-CNN (Dong et al. 2015) network proposed by Dong et al. in 2015, which can reduce compression artifacts with only three convolutional layers. Due to the limitations of shallow networks in extracting image features, Zhang et al. developed a deeper DnCNN (Zhang et al. 2017) using residual learning, which demonstrated strong performance in image denoising tasks. Although DnCNN demonstrates better performance in removing blind noise, its deeper network layers lead to increased memory consumption. To further improve the quality of compressed videos, Wang et al. proposed DCAD (Wang, Chen, and Chao 2017), which stacks convolutional layers to enhance the model's receptive field and effectively mitigate distortion issues. It is known that video compression contains intra-coded (I-frames) and inter-coded (P/B-frames) modes; these different encoding modes cause varying degrees of distortion in the compressed frames. To address this, Yang et al. proposed QE-CNN (Yang et al. 2018), which employs two specialized subnetworks, QE-CNN-I and QE-CNN-P, to separately optimize I-frames and P-frames. However, these methods neglect the spatiotemporal correlations among multiple frames and cannot leverage information from neighboring frames, severely limiting video reconstruction performance. Later, multi-frame enhancement-based approaches became mainstream, significantly improving video restora-

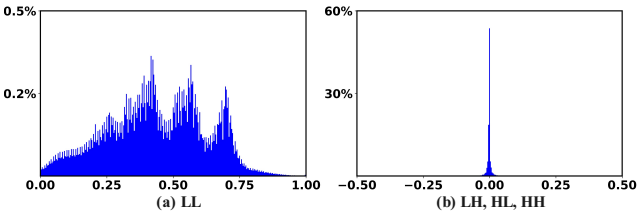


Figure 1: Histogram of the distribution (value-probabilities) of BasketballPass in the frequency domain.

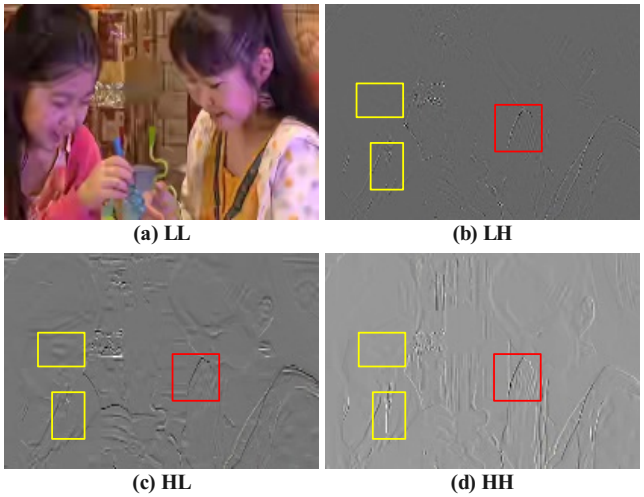


Figure 2: Visualization of the four sub-bands obtained after wavelet transform.

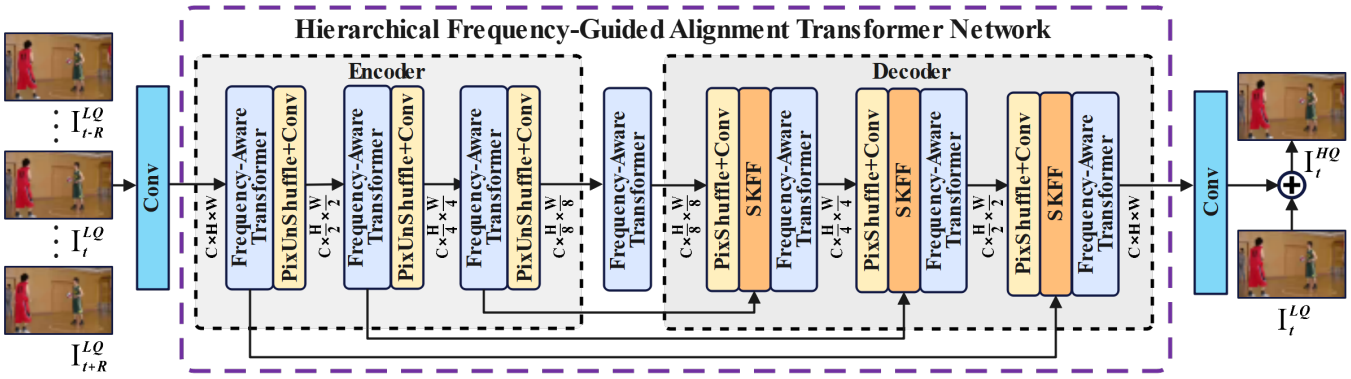


Figure 3: Architecture of our Hierarchical Frequency-Guided Alignment Transformer network.

tion performance.

## Multi-Frame-based Video Enhancement

In the early stage, Tran et al. employed 3D convolutions (Tran et al. 2015) to directly learn the spatio-temporal correlations among multiple frames. However, constrained by the influence of object motion, this early fusion scheme struggled to effectively utilize the information from adjacent frames. To address this issue, Guan et al. proposed a motion compensation network (Guan et al. 2019) based on optical flow. This network utilized optical flow for multi-frame alignment, thereby achieving effective information fusion. Nevertheless, due to the presence of compression artifacts, the accuracy of optical flow prediction was compromised, leading to the current frame learning some irrelevant information from adjacent frames. Subsequently, Deng et al. introduced a spatio-temporal deformable fusion strategy (STDF) (Deng et al. 2020) that employed deformable convolutions for multi-frame alignment. Given the limitations of single-scale alignment in the spatio-temporal deformable fusion strategy, Luo et al. introduced a multi-path deformable alignment module (Luo et al. 2022). By integrating alignment features with different receptive fields, this module more effectively utilized the information from adjacent frames. To further expand the model's receptive field, Yu et al. utilized a Transformer (Yu et al. 2024) to learn the spatio-temporal dependencies among multiple frames. Since compressed videos are primarily affected by frequency-domain distortions, these spatial-domain-based feature fusion methods struggle to effectively capture the complementary frequency information between adjacent frames, leading to insufficient reconstruction of high-frequency details and inaccurate recovery of low-frequency structures.

# Proposed Method

Given a compressed video sequence $\{I_1^{LQ}, I_2^{LQ}, ..., I_N^{LQ}\}$ consisting of N low-quality (LQ) frames, where $I_t^{LQ}$ denotes the compressed frame at time $t$. We use a set of 2R+1 neighboring frames, $I_t^{2R+1} = \{I_{t-R}^{LQ}, I_{t-R+1}^{LQ}, ..., I_{t+R}^{LQ}\}$, centered around $I_t^{LQ}$, as input to a quality enhancement model, and obtain an enhanced frame $I_t^{HQ}$. By iteratively processing N times, we achieve a progressive restoration of the quality of the entire video. This process can be formulated as:

$$I_t^{HQ} = F(I_t^{2R+1}, \theta), t = 1, 2, \ldots, N \quad (1)$$

where $F$ is the Hierarchical Frequency-Guided Alignment Transformer network, and $\theta$ is the learnable parameter.

As shown in Fig.3, the proposed framework adopts a hierarchical structure (encoder-decoder). During the encoding stage, we utilize the PixelUnShuffle module to progressively downsample features, obtaining multi-scale features (spatial), and enhance these features using Frequency-Aware Transformer modules. In the decoding stage, features are progressively upsampled via the PixelShuffle module to reconstruct high-resolution features, while additional Frequency-Aware Transformer modules are introduced to further refine the upsampled features. To ensure the consistency of the channel dimensions between encoder and decoder features, convolutional layers are added after both the PixelUnShuffle and PixelShuffle modules to adjust the number of channels. Furthermore, we employ the SKFF (Zamir et al. 2022) module to adaptively fuse the features from the two branches. Specifically, the SKFF module leverages an attention mechanism to assign adaptive weights, denoted as $s1$ and $s2$, to the features $L1$ and $L2$, thereby adjusting the fusion process as follows: $L = L1 \times s1 + L2 \times s2$.

## Frequency-Aware Transformer Module

As shown in Figure 4, the Frequency-Aware Transformer module consists of four primary components: the Frequency Decomposition block, the Frequency-Independent Transformer block, the Frequency-Mixed Transformer block, and the Inverse Synthesis block. The Frequency Decomposition block transforms spatial domain signals into their corresponding frequency domain representations. The Frequency-Independent Transformer block performs self-attention separately within each sub-band, thereby avoiding information interference between different sub-bands. The Frequency-Mixed Transformer block implements cross-band self-attention, capable of deeply exploring the similar information contained between different sub-bands. Finally, the inverse synthesis block transforms frequency-domain features into spatial-domain representations, completing the frequency-to-spatial conversion.

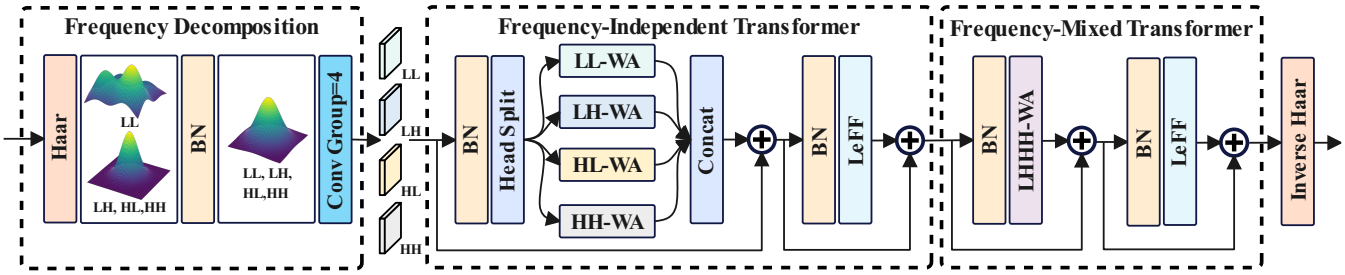


Figure 4: Architecture of our Frequency-Aware Transformer block.

**Frequency Decomposition Block and Inverse Synthesis Block.** First, we apply the Haar wavelet transform to convert the spatial signal into the frequency domain. To better accommodate the significant distribution differences between high-frequency and low-frequency components after wavelet transform, we introduce a normalization layer after frequency domain decomposition. By adjusting the mean and variance of features across different frequencies, the normalization layer reduces the distribution gap among them, thereby preventing the model from overemphasizing certain frequency components and neglecting other crucial information during training. Although Layer Normalization (LN) can partially mitigate the distribution discrepancy between low- and high-frequency features, it forces all channels (i.e., LL, LH, HL, and HH) to share identical statistical distributions. However, different channels contain distinct semantic and structural information, and this forced distribution sharing may disrupt the structural features of different channels, thereby potentially reducing model performance. In contrast, Batch Normalization (BN) normalizes each channel independently across a batch of samples, thereby better preserving the distinct structural features of each channel, which may offer greater advantages for fusing information from adjacent frames in the frequency domain. To further extract features from the frequency domain, a convolutional layer with a group size of four is employed:

$$f_{freq} = Conv(BN(Haar(f_{spatial})), group = 4). \quad (2)$$

where $Haar$ denotes the Haar wavelet transform. At the last layer of the Frequency-Aware Transformer module, we convert wavelet domain features into spatial domain features through the inverse Haar wavelet transform:

$$f_{spatial} = Conv(InverseHaar(f_{freq})). \quad (3)$$

**Frequency-Independent Transformer Block.** The Frequency-Independent Transformer consists of three core components: a BN layer for feature standardization; a frequency-independent self-attention mechanism for modeling intra-subband correlations; and a Locally Enhanced Feed-Forward Network (LeFF) (Wang et al. 2022) for capturing local context. The frequency-independent self-attention mechanism employs a four-head (heads=4) architecture, dividing the input features along the channel dimension into four frequency subbands (LL, LH, HL, HH), with independent self-attention computations within each subband. This design effectively captures the intra-subband correlations while efficiently preventing interference between different subbands. Subsequently, the LeFF module further explores the local contextual information within the frequency domain:

$$\hat{f}_{freq} = WA_{LL,LH,HL,HH}(BN(f_{freq})) + f_{freq}, \quad (4)$$

$$\tilde{f}_{freq} = LeFF(BN(\hat{f}_{freq})) + \hat{f}_{freq}. \quad (5)$$

where $WA_{LL,LH,HL,HH}$ is a multi-head self-attention module that performs self-attention separately in each frequency domain. This process can be formulated as:

$$\widehat{f}_{freq} = \{f_{LL}, f_{LH}, f_{HL}, f_{HH}\}, \quad (6)$$

$$Q_i, K_i, V_i = f_i W_i^{QKV}, i = LL, LH, HL, HH, \quad (7)$$

$$Attention(Q, K, V) = \text{SoftMax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i. \quad (8)$$

where $W_i^{QKV}$ denotes the projection matrices for the $i-th$ frequency subband, which linearly map the input features $f_i$ to the queries, keys, and values, respectively. $d_k$ is the channel number of projected matrices.

**Frequency-Mixed Transformer Block.** The Frequency-Mixed Transformer block includes a BN layer, a frequency-mixed self-attention mechanism, and a LeFF module. In the implementation of the frequency-mixed self-attention, we utilize a single-head attention mechanism that jointly models all frequency sub-bands (LL, LH, HL, HH) using a unified attention head. This approach explicitly captures cross-band dependencies, allowing the current sub-band to aggregate contextual information from other sub-bands, thereby enabling comprehensive feature interaction and information fusion.

## Training Scheme

We employ a joint optimization strategy that integrates Mean Squared Error (MSE) with Learned Perceptual Image Patch Similarity (LPIPS) loss for training. This combination effectively mitigates the over-smoothing tendency of MSE and reduces artifacts and noise associated with sole reliance on LPIPS, leading to improved objective and perceptual quality. The overall loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{mse}(I_t^{HQ}, I_t^{GT}) + \lambda \times \mathcal{L}_{lpips}(I_t^{HQ}, I_t^{GT}) \quad (9)$$

where $I_t^{HQ}$ denotes the reconstructed frame, and $I_t^{GT}$ represents the corresponding ground-truth frame. $\lambda$ is a hyper-parameter used to balance objective quality and perceptual quality.

| QP | Class | Sequence | Dong et al. | | | STCF | | | TGAF | | | HFUR | | | Our | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS |
| 42 | A | Traffic | - | - | -5.50 | -0.20 | -1.553 | -8.09 | -0.12 | -1.245 | -5.47 | -0.23 | -0.790 | -4.47 | **0.18** | **0.212** | **-8.56** |
| | | PeopleOnStreet | - | - | -1.40 | 0.30 | 0.580 | -2.12 | 0.42 | 0.760 | -3.07 | 0.43 | 1.490 | 1.54 | **0.60** | **1.683** | **-5.20** |
| | B | Kimono | - | - | -2.90 | -0.28 | -2.624 | -1.75 | -0.19 | -2.464 | -0.46 | -0.13 | **-0.728** | -1.41 | **-0.03** | -1.499 | **-4.31** |
| | | ParkScene | - | - | -11.90 | -0.13 | -1.424 | **-14.68** | -0.12 | -1.885 | -9.75 | -0.12 | **-0.356** | -8.00 | **0.04** | -0.725 | -13.20 |
| | | Cactus | - | - | **-10.70** | -0.12 | -0.958 | -9.14 | -0.05 | -0.796 | -7.57 | -0.14 | -0.260 | -3.96 | **0.21** | **0.157** | -10.17 |
| | | BQTerrace | - | - | -5.70 | -0.02 | -0.999 | -6.72 | 0.08 | -0.431 | -5.89 | 0.04 | 0.129 | -3.90 | **0.28** | **0.333** | **-8.53** |
| | | BasketballDrive | - | - | -10.20 | 0.02 | -0.962 | -8.69 | 0.04 | -1.212 | -9.18 | 0.14 | **0.131** | -6.13 | **0.28** | 0.097 | **-11.82** |
| | C | RaceHorses | - | - | **-7.20** | 0.09 | -0.619 | -6.36 | 0.01 | -0.926 | -4.98 | 0.08 | **0.401** | -3.01 | **0.20** | 0.026 | -5.95 |
| | | BQMall | - | - | -7.20 | 0.16 | -0.109 | -6.63 | 0.17 | -0.231 | -6.38 | 0.20 | 0.594 | -7.00 | **0.40** | **0.842** | **-8.09** |
| | | PartyScene | - | - | -8.50 | -0.06 | -0.788 | -7.72 | -0.07 | -0.769 | -6.10 | -0.03 | 0.724 | -6.41 | **0.30** | **1.943** | **-8.59** |
| | | BasketballDrill | - | - | -5.60 | -0.01 | -1.344 | -6.75 | -0.02 | -1.503 | -7.15 | 0.02 | 0.541 | -2.24 | **0.36** | **0.992** | **-7.80** |
| | D | RaceHorses | - | - | -5.40 | 0.20 | -0.005 | **-6.59** | 0.13 | -0.377 | -6.54 | 0.31 | **1.314** | -4.56 | **0.33** | 0.778 | -6.53 |
| | | BQSquare | - | - | -4.40 | 0.00 | -0.473 | -5.46 | 0.08 | -0.331 | -4.83 | 0.16 | **0.482** | -4.33 | **0.25** | 0.462 | **-7.59** |
| | | BlowingBubbles | - | - | -9.40 | -0.03 | -0.887 | -9.85 | -0.02 | -0.784 | -7.90 | 0.06 | 0.353 | -8.42 | **0.33** | **1.603** | **-10.73** |
| | | BasketballPass | - | - | -6.00 | 0.21 | -0.199 | **-7.38** | 0.19 | -0.437 | -4.83 | 0.34 | 1.018 | -5.74 | **0.44** | **1.120** | -6.93 |
| | E | FourPeople | - | - | -3.80 | -0.18 | -1.011 | -3.35 | -0.17 | -1.122 | -3.38 | -0.37 | -0.690 | -2.37 | **0.33** | **0.211** | **-6.09** |
| | | Johnny | - | - | -6.20 | -0.35 | -1.750 | -8.54 | -0.27 | -1.784 | -7.67 | -0.79 | -1.319 | -3.37 | **0.18** | **-0.494** | **-8.79** |
| | | KristenAndSara | - | - | -4.40 | -0.11 | -0.774 | -6.42 | -0.19 | -1.392 | -4.93 | -0.63 | -1.022 | -1.89 | **0.37** | **0.009** | **-7.30** |
| | **Average** | | - | - | -6.50 | -0.03 | -0.883 | -7.01 | 0.00 | -0.940 | -5.89 | -0.03 | 0.112 | -4.20 | **0.28** | **0.431** | **-8.12** |
| **37** | **Average** | | - | - | -6.40 | 0.06 | -0.420 | -5.75 | -0.02 | -1.153 | -5.28 | 0.13 | **0.125** | -3.87 | **0.24** | 0.019 | **-7.24** |

Table 1: Overall comparison for $\Delta$PSNR (dB), $\Delta$LPIPS ($\times 10^{-2}$) and $\Delta$SSIM ($\times 10^{-2}$) over test sequences.

# Experiments

## Dataset

To comprehensively evaluate our model's performance, we adopt the same experimental setup as RDFN (Peng et al. 2022a) and STCF (Zhang et al. 2023). The datasets are sourced from Xiph.org[1], VQEG[2], and JCT-VC (Bossen 2010), comprising a total of 126 video sequences. Consistent with the dataset partitioning strategy of the STCF, 108 sequences are used for training, while the remaining 18 are reserved for evaluation. Both the training and test sets cover videos of varying resolutions and diverse content types. All videos are encoded using the H.265/HEVC reference software HM16.5 and the H.266/VVC reference software VTM17.0 in low-delay (LDP) mode.

## Implementation Details

In our experiments, all models were trained on a server equipped with eight Tesla V100-SXM2-16GB GPUs. During training, we randomly cropped 128×128 patches from the original videos and their corresponding compressed videos to serve as training samples, with a batch size of 32. All models were optimized using the Adam optimizer (Kingma and Ba 2014) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-6}$), with an initial learning rate of $1 \times 10^{-4}$, and trained for a total of 400 epochs. During model training, different $\lambda$ values were used for different datasets: 0.01 for HM16.5 and 0.005 for VTM17.0 compressed datasets. Consistent with previous methods based on perceptual quality enhancement (Chen et al. 2025; Wang et al. 2021, 2020), we opt to evaluate the model performance in the RGB color space. Beyond perceptual quality ($\Delta$LPIPS) and subjective quality, we also conduct a comprehensive evaluation of objective indicators ($\Delta$PSNR/$\Delta$SSIM) to further substantiate the effectiveness and superiority of our approach.

## Comparison With State-of-the-Art Methods

To ensure a fair and comprehensive evaluation, in the experimental section, we will conduct an all-around comparison between our proposed method and the state-of-the-art compressed video enhancement methods developed in recent years, including Dong et al. (Dong et al. 2024), STDR (Luo et al. 2022), STCF (Zhang et al. 2023), TGAF (Zhu et al. 2024b), and HFUR (Zhang et al. 2025b). Among these, the method proposed by Dong et al. represents the state-of-the-art approach for perceptual quality enhancement, while HFUR stands as the leading-edge technique for objective quality enhancement.

**Quantitative Results.** Tables 1 and 2, respectively, present the enhancement effects on videos compressed by HM16.5 and VTM17.0. The results demonstrate that our proposed method outperforms all other comparative approaches in terms of the PSNR metric across the board. Although the LPIPS metric of our method is slightly inferior to that of certain other methods on some videos, its average LPIPS across all videos is higher than that of the comparison methods. Even though under the condition of QP=37, our method exhibits a slightly lower SSIM compared to HFUR, it achieves significant improvements in both PSNR and LPIPS. Moreover, by adjusting the $\lambda$, it is entirely feasible to bridge the gap in SSIM. Specifically, as shown in Table 1, at QP = 42, our method reduces the LPIPS metric by 0.0162 compared to the state-of-the-art perceptual quality enhancement approach proposed by Dong et al. Addition-

[1]https://media.xiph.org/video/derf

[2]https://vqeg.org/video-datasets-and-organizations.aspx

| QP | Class | Sequence | STDR | | | STCF | | | TGAF | | | HFUR | | | Our | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS | PSNR | SSIM | LPIPS |
| 42 | A | Traffic | -0.38 | -2.517 | -9.11 | -0.34 | -2.585 | -10.42 | -0.20 | -1.778 | -8.44 | -0.28 | -0.933 | -8.13 | **0.01** | **-0.749** | **-12.05** |
| | | PeopleOnStreet | 0.05 | -0.692 | -6.22 | 0.01 | -1.127 | -6.66 | 0.16 | -0.424 | -6.99 | 0.11 | 0.052 | -3.18 | **0.33** | **0.313** | **-7.54** |
| | B | Kimono | -0.22 | -2.423 | -10.60 | -0.17 | -2.107 | -9.42 | -0.07 | -1.792 | -8.96 | -0.15 | -0.838 | -8.92 | **0.08** | **-0.752** | **-13.44** |
| | | ParkScene | -0.28 | -2.888 | -20.26 | -0.14 | -1.477 | **-20.47** | -0.10 | -1.664 | -16.92 | -0.14 | **-0.462** | -13.22 | **0.05** | -0.603 | -19.14 |
| | | Cactus | -0.40 | -2.357 | -10.46 | -0.24 | -1.614 | -11.54 | -0.19 | -1.339 | -10.10 | -0.32 | -1.117 | -5.46 | **0.04** | **-0.523** | **-12.49** |
| | | BQTerrace | -0.35 | -2.069 | -11.19 | -0.36 | -2.087 | -10.11 | -0.17 | -1.168 | -8.94 | -0.23 | **-0.650** | -5.65 | **-0.09** | -0.780 | **-11.46** |
| | | BasketballDrive | -0.43 | -2.602 | -10.02 | -0.28 | -1.809 | -9.60 | -0.20 | -2.149 | -10.66 | -0.22 | **-0.642** | -5.23 | **0.02** | -1.070 | **-13.78** |
| | C | RaceHorses | -0.13 | -1.639 | -8.44 | -0.07 | -1.296 | -8.61 | -0.11 | -1.506 | -8.58 | -0.06 | **-0.565** | -5.52 | **0.02** | -0.981 | **-10.19** |
| | | BQMall | -0.08 | -1.683 | -8.63 | -0.03 | -1.301 | -8.48 | 0.02 | -1.145 | -7.56 | 0.08 | **-0.342** | -8.02 | **0.19** | -0.424 | **-10.25** |
| | | PartyScene | -0.28 | -2.077 | -6.90 | -0.32 | -3.153 | -6.64 | -0.24 | -2.338 | -5.64 | -0.13 | -0.886 | -6.26 | **0.12** | **-0.113** | **-8.99** |
| | | BasketballDrill | -0.51 | -2.759 | -8.18 | -0.38 | -2.237 | -6.71 | -0.26 | -2.395 | -7.88 | -0.35 | -1.002 | -2.55 | **0.01** | **-0.795** | **-10.26** |
| | D | RaceHorses | 0.02 | -1.034 | -9.63 | 0.08 | -0.607 | -7.67 | 0.02 | -0.970 | -9.46 | 0.14 | **0.087** | -7.82 | **0.19** | -0.213 | **-10.69** |
| | | BQSquare | -0.26 | -1.322 | -6.43 | -0.22 | -1.597 | -7.34 | -0.17 | -1.134 | -5.13 | -0.06 | -0.940 | -5.01 | **0.05** | **-0.396** | **-7.42** |
| | | BlowingBubbles | -0.16 | -1.640 | -9.87 | -0.19 | -2.305 | -10.63 | -0.11 | -1.663 | -8.33 | -0.02 | -0.692 | -9.76 | **0.22** | **0.478** | **-12.56** |
| | | BasketballPass | -0.02 | -1.316 | -7.63 | 0.09 | -0.686 | -8.88 | 0.05 | -1.284 | -6.80 | 0.14 | **0.239** | -7.14 | **0.31** | 0.138 | **-9.80** |
| | E | FourPeople | -0.49 | -2.398 | -4.87 | -0.47 | -2.113 | -3.77 | -0.30 | -1.821 | -4.63 | -0.46 | -1.419 | -3.57 | **0.08** | **-0.707** | **-7.21** |
| | | Johnny | -1.08 | -2.280 | **-8.41** | -0.60 | -2.303 | -8.26 | -0.47 | -2.171 | -8.07 | -0.81 | -1.423 | -4.14 | **-0.06** | **-0.801** | -7.79 |
| | | KristenAndSara | -0.77 | -1.923 | -7.61 | -0.50 | -1.610 | -7.64 | -0.44 | -1.871 | -7.29 | -0.77 | -1.417 | -3.92 | **0.04** | **-0.422** | **-8.20** |
| | **Average** | | -0.32 | -1.979 | -9.14 | -0.23 | -1.779 | -9.05 | -0.15 | -1.589 | -8.35 | -0.20 | -0.719 | -6.31 | **0.09** | **-0.467** | **-10.74** |
| **37** | **Average** | | -0.32 | -1.400 | -8.24 | -0.18 | -1.040 | -8.31 | -0.13 | -1.112 | -7.31 | -0.26 | -0.660 | -5.82 | **0.07** | **-0.291** | **-8.88** |

Table 2: Overall comparison for $\Delta$PSNR (dB), $\Delta$LPIPS ($\times 10^{-2}$) and $\Delta$SSIM ($\times 10^{-2}$) over test sequences.

ally, relative to the advanced objective quality enhancement method HFUR, our approach achieves an increase of 0.31 dB in PSNR and a gain of 0.00319 in SSIM. Moreover, when the hyperparameter $\lambda$ is set to 0.01, our proposed method not only achieves a significant reduction in the LPIPS value but also leads to improvements in the PSNR and SSIM metrics.

**Qualitative Results.** Figure 5 presents a qualitative comparison of enhancement results across different methods. It can be observed that the compression process inevitably results in the loss of high-frequency information and the appearance of false textures. Compared to other state-of-the-art approaches, our method demonstrates a significant advantage in restoring lost details, effectively reconstructing authentic fine structures while simultaneously removing spurious high-frequency textures. These results further validate the superior performance of the proposed method in restoring video subjective quality.

**Model Complexity.** Similar to previous methods (Peng et al. 2022b; Zhao, Xu, and Zhou 2021), we evaluate model complexity using the model parameters, FLOPs (Floating-Point Operations), and FPS (Frames Per Second). Table 3 compares the model complexity across different approaches. It can be seen that our method achieves superior performance in both computational efficiency and inference speed, while also maintaining a competitive number of parameters. Specifically, compared to the state-of-the-art HFUR method, our approach requires only one-twentieth of its FLOPs and achieves eleven times higher FPS. Compared to STCF, our method's FLOPs are reduced to one-third, with FPS tripling that of STCF. The results demonstrate that our method achieves significantly lower complexity compared to other state-of-the-art approaches, further validating its superiority.

| Method | Parameter(M) | FLOPs(G) | FPS(720p) | $\Delta$LPIPS($\times 10^{-2}$) |
|---|---|---|---|---|
| STDR | **1.48** | 859.23 | 5.3 | -9.14 |
| STCF | 2.08 | 601.86 | 2.1 | -9.05 |
| TGAF | 1.78 | 1667.63 | 4.6 | -8.35 |
| HFUR | 6.52 | 6749.59 | 0.6 | -6.31 |
| Ours | 1.70 | **205.23** | **6.9** | **-10.74** |

Table 3: Comparison of model complexity, with FLOPs and FPS evaluated on Class E.

| Method | $\Delta$PSNR(dB) | $\Delta$SSIM($\times 10^{-2}$) | $\Delta$LPIPS($\times 10^{-2}$) |
|---|---|---|---|
| Spatial | 0.05 | -0.914 | -7.17 |
| Frequency | **0.28** | **0.431** | **-8.12** |

Table 4: Ablation Studies on different alignment methods.

## Ablation Study

**Effect of Frequency-Domain Alignment.** To evaluate the effectiveness of frequency-domain alignment, we trained a similar complexity spatial-domain alignment Transformer. As shown in Table 4, the frequency-domain alignment model significantly outperforms the spatial-domain model in all indicators. Specifically, the frequency-domain approach achieves an improvement of 0.23 dB in PSNR and a reduction of 0.0095 in LPIPS metric, thereby providing strong validation for the effectiveness of the frequency-domain alignment method.

**Effect of combining frequency-independent processing and frequency-mixed processing.** To validate the effectiveness of the proposed strategy that combines frequency-independent processing with frequency-mixed process-

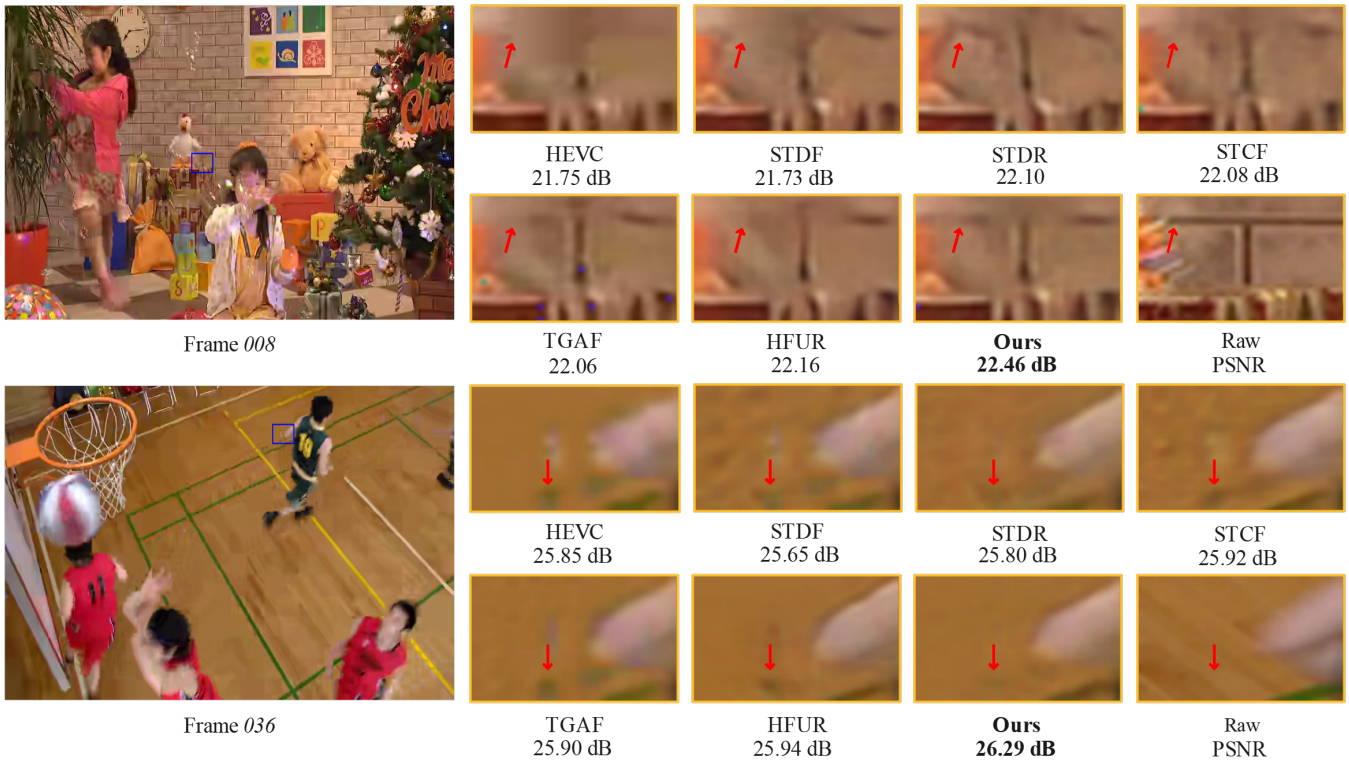


Figure 5: The qualitative comparison results shown in the images for PartyScene (top) and BasketballDrill (bottom) videos indicate that our method performs better in restoring structural details.

ing, we conducted ablation experiments by replacing the frequency-mixed self-attention with frequency-independent self-attention, denoted as (only FIT). Conversely, we also replaced the frequency-independent self-attention with frequency-mixed self-attention, denoted as (only FMT). In the full model, we adopted a scheme that integrates both frequency-independent and mixed processing (FIT+FMT). As shown in Table 5, compared to the method utilizing only frequency-mixed self-attention, the combined strategy achieved an improvement of 0.08 dB in PSNR and 0.0042 in SSIM. Furthermore, when compared to the pure frequency-independent processing scheme, this method demonstrated an increase of 0.04 dB in PSNR, along with a reduction of 0.0027 in the LPIPS metric. The experimental results demonstrate that integrating frequency-independent processing with mixing processing can more effectively facilitate information exchange between different sub-bands and reduce cross-interference from irrelevant information among adjacent sub-bands.

**Effect of Batch Normalization.** To assess the impact of the BN in the frequency domain, we conducted ablation experiments by removing the BN layers within the frequency decomposition block and replacing the BN layers in both the frequency-independent and frequency-mixed self-attention modules with LN. As shown in Table 5, the model incorporating BN consistently achieves superior performance in PSNR, SSIM, and LPIPS metrics, indicating that BN is more effective for normalizing frequency-domain features.

| Method | $\Delta$PSNR(dB) | $\Delta$SSIM($\times 10^{-2}$) | $\Delta$LPIPS($\times 10^{-2}$) |
|---|---|---|---|
| only FMT+LN | 0.15 | -0.339 | -8.04 |
| only FIT+LN | 0.19 | 0.131 | -7.75 |
| FIT+FMT+LN | 0.23 | 0.091 | -8.02 |
| FIT+FMT+BN | **0.28** | **0.431** | **-8.12** |

Table 5: Ablation Studies on different modules.

## Conclusion

This paper first analyzes the characteristics in the frequency domain, revealing that different sub-bands are both interrelated and possess a certain degree of independence. Based on these findings, we propose a frequency-aware transformer module that further enhances information interaction across different frequency domains by combining independent frequency-domain processing with hybrid frequency-domain processing. Additionally, we observe significant differences in the data distributions of different sub-bands. Therefore, after frequency-domain decomposition, we employ batch normalization instead of layer normalization to preserve the structural features of distinct sub-bands while alleviating training difficulties caused by distribution disparities. Experimental results demonstrate that our method achieves state-of-the-art performance in terms of objective metrics (PSNR/SSIM), perceptual quality (LPIPS), and subjective visual effects, while also significantly reducing model complexity.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant U24B20132, 62271290 and 62572275, and in part by the Shandong Provincial Natural Science Foundation under Grant ZR2024LZN021, ZR2025QB44, ZR2025MS1074 and ZR2022ZD38.

## References

Bossen, F. 2010. Common test conditions and software reference configurations. In *3rd. JCT-VC Meeting, Guangzhou, CN, October 2010*.

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, J.; Chen, K.; Zeng, H.; Lin, Q.; and Zhu, J. 2025. Perceptual Quality Enhancement for Compressed Video With High-Frequency Details and High-Dimensional Features. *IEEE Transactions on Instrumentation and Measurement*, 74: 1–13.

Deng, J.; Dong, S.; Chen, L.; Hu, J.; and Zhuo, C. 2024. Stdf: Spatio-temporal deformable fusion for video quality enhancement on embedded platforms. *ACM Transactions on Embedded Computing Systems*, 23(2): 1–25.

Deng, J.; Wang, L.; Pu, S.; and Zhuo, C. 2020. Spatio-temporal deformable convolution for compressed video quality enhancement. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 10696–10703.

Ding, Q.; Shen, L.; Yu, L.; Yang, H.; and Xu, M. 2021. Patch-wise spatial-temporal quality enhancement for HEVC compressed video. *IEEE Transactions on Image Processing*, 30: 6459–6472.

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.

Dong, C.; Ma, H.; Li, Z.; Li, L.; and Liu, D. 2024. Temporal Wavelet Transform-Based Low-Complexity Perceptual Quality Enhancement of Compressed Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5): 4040–4053.

Forecast, G.; et al. 2019. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022. *Update*, 2017: 2022.

Guan, Z.; Xing, Q.; Xu, M.; Yang, R.; Liu, T.; and Wang, Z. 2019. MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video. *IEEE transactions on pattern analysis and machine intelligence*, 43(3): 949–963.

He, G.; Quan, G.; Wu, C.; Wang, S.; Zhou, D.; and Li, Y. 2025. Multi-Frame Deformable Look-Up Table for Compressed Video Quality Enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3392–3400.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Liu, J.; Zhou, M.; and Xiao, M. 2022. Deformable convolution dense network for compressed video quality enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1930–1934. IEEE.

Luo, D.; Ye, M.; Chen, S.; and Li, X. 2021. Alignment-free video compression artifact reduction. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, 1–5. IEEE.

Luo, D.; Ye, M.; Li, S.; Zhu, C.; and Li, X. 2022. Spatio-temporal detail information retrieval for compressed video quality enhancement. *IEEE Transactions on Multimedia*, 25: 6808–6820.

Peng, L.; Hamdulla, A.; Ye, M.; Li, S.; and Guo, H. 2022a. Recurrent deformable fusion for compressed video artifact reduction. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, 3175–3179. IEEE.

Peng, L.; Hamdulla, A.; Ye, M.; Li, S.; Wang, Z.; and Li, X. 2022b. OVQE: Omniscient network for compressed video quality enhancement. *IEEE Transactions on Broadcasting*, 69(1): 153–164.

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.

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

Wang, J.; Deng, X.; Xu, M.; Chen, C.; and Song, Y. 2020. Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video. In *European conference on computer vision*, 405–421. Springer.

Wang, J.; Xu, M.; Deng, X.; Shen, L.; and Song, Y. 2021. MW-GAN+ for perceptual quality enhancement on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4224–4237.

Wang, M.; Liao, Y.; Chen, W.; Lin, L.; and Zhao, T. 2025. STFF: Spatio-Temporal and Frequency Fusion for Video Compression Artifact Removal. *IEEE Transactions on Broadcasting*, 71(2): 542–554.

Wang, T.; Chen, M.; and Chao, H. 2017. A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC. In *2017 data compression conference (DCC)*, 410–419. IEEE.

Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; and Li, H. 2022. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17683–17693.

Xu, Y.; Zhao, M.; Liu, J.; Zhang, X.; Gao, L.; Zhou, S.; and Sun, H. 2021. Boosting the performance of video compression artifact reduction with reference frame proposals and frequency domain information. In *Proceedings of*

*the IEEE/CVF conference on computer vision and pattern recognition*, 213–222.

Yang, R.; Xu, M.; Liu, T.; Wang, Z.; and Guan, Z. 2018. Enhancing quality for HEVC compressed videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(7): 2039–2054.

Yu, L.; Chang, W.; Wu, S.; and Gabbouj, M. 2024. End-to-End Transformer for Compressed Video Quality Enhancement. *IEEE Transactions on Broadcasting*, 70(1): 197–207.

Yue, J.; Gao, Y.; Li, S.; Yuan, H.; and Dufaux, F. 2022. A global appearance and local coding distortion based fusion framework for CNN based filtering in video coding. *IEEE Transactions on Broadcasting*, 68(2): 370–382.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; Yang, M.-H.; and Shao, L. 2022. Learning enriched features for fast image restoration and enhancement. *IEEE transactions on pattern analysis and machine intelligence*, 45(2): 1934–1948.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7): 3142–3155.

Zhang, M.; Bai, H.; Shang, W.; Guo, J.; Li, Y.; and Gao, X. 2025a. MDEformer: Mixed Difference Equation Inspired Transformer for Compressed Video Quality Enhancement. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 2410–2422.

Zhang, Q.; Zheng, B.; Chen, X.; Chen, Q.; Zhu, Z.; Wang, C.; Li, Z.; Jia, X.; and Yan, C. 2025b. Hierarchical Frequency-Based Upsampling and Refining for HEVC Compressed Video Enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5): 4423–4436.

Zhang, S.; Herranz, L.; Mrak, M.; Blanch, M. G.; Wan, S.; and Yang, F. 2022. DCNGAN: A Deformable Convolution-Based GAN with QP Adaptation for Perceptual Quality Enhancement of Compressed Video. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2035–2039.

Zhang, T.; He, X.; Teng, Q.; Cheng, J.; and Ren, C. 2024. Spatio-Temporal Adaptive Weighted Fusion Network for Compressed Video Quality Enhancement. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 71(12): 5064–5068.

Zhang, X.; Yang, S.; Luo, W.; Gao, L.; and Zhang, W. 2023. Video compression artifact reduction by fusing motion compensation and global context in a swin-CNN based parallel architecture. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3489–3497.

Zhao, M.; Xu, Y.; and Zhou, S. 2021. Recursive fusion and deformable spatiotemporal attention for video compression artifact reduction. In *Proceedings of the 29th ACM international conference on multimedia*, 5646–5654.

Zhu, Q.; Hao, J.; Ding, Y.; Liu, Y.; Mo, Q.; Sun, M.; Zhou, C.; and Zhu, S. 2024a. CPGA: Coding priors-guided aggregation network for compressed video quality enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2964–2974.

Zhu, Q.; Qiu, Y.; Liu, Y.; Zhu, S.; and Zeng, B. 2024b. Compressed video quality enhancement with temporal group alignment and fusion. *IEEE Signal Processing Letters*, 31: 1565–1569.