Temporal Wavelet Transform-Based Low-Complexity Perceptual Quality Enhancement of Compressed Video

Cunhui Dong, Haichuan Ma, Zhuoyuan Li, *Student Member, IEEE*, Li Li, *Member, IEEE*, and Dong Liu, *Senior Member, IEEE*

***Abstract*— The past few years have witnessed a great success in applying deep learning to enhance the perceptual quality of compressed video. These methods usually perform frame-by-frame quality enhancement, incurring high computational complexity. Low-complexity perceptual quality enhancement is addressed in this paper, motivated by the observation of temporal correlations among video frames. We propose to decompose video content into *temporal* low-frequency and high-frequency components, and to focus the enhancement of the temporal low-frequency component, which may significantly reduce the computational complexity. Specifically, we employ the temporal wavelet transform (TWT) for the temporal frequency analysis, and build a TWT-based multiple-input multiple-output perceptual quality enhancement scheme. First, we use a motion estimation method on the input video to acquire the motion information, and then use TWT to obtain the temporal low- and high-frequency components. Second, we design a deep network to enhance the quality of the temporal low-frequency component. Finally, the temporal high-frequency component and the enhanced temporal low-frequency component are combined by the temporal wavelet inverse transform (TWIT) to generate the enhanced video. Experimental results show that our method achieves comparable perceptual quality to that of the state-of-the-art methods, but reduces the computational complexity to 1/13.**

***Index Terms*— Deep learning, low complexity, perceptual quality enhancement, temporal wavelet transform, video compression.**

I. Introduction

NOWADAYS, with the dramatic growth of data traffic over the internet and the emergent application of versatile video formats such as 2K, 4K, high dynamic range, and wide color gamut, there is a pressing demand for storage and transmission. To address this challenge, in recent decades, lossy compression is employed to significantly reduce the amount of data. However, lossy compression inevitably causes compression artifacts (such as blurring and ringing effects) and significantly compromises the quality of experience (QoE) for viewers. To solve this problem, quality enhancement on compressed videos has been widely studied in recent years to mitigate the compression artifacts and improve the QoE.

In recent years, with the rapid evolution of deep neural networks (DNN), many DNN-based methods have been proposed to enhance the visual quality of compressed images/videos. References [1], [2], [3], and [4] focus on the objective quality enhancement for compressed images/videos. However, it is found that improving the objective quality may not correlate with the enhancement of perceptual quality in real application [5]. To improve the QoE, many methods have been tentatively proposed to optimize the perceptual quality. Galteri *et al*. [6], [7] design a generative adversarial network (GAN) for image compression artifact reduction, which realistically recovers high-frequency details. CVRGAN [8] firstly introduces the GAN loss in network training strategy to enhance the perceptual quality of compressed videos. MW-GAN [9] proves that spatial high-frequency has a great influence on the perceptual quality, so the wavelet packet is employed to decompose the spatial high- and low-frequency content in frames. Based on MW-GAN, MW-GAN+ [10] further introduces an advanced motion alignment network and a 3D discriminator to improve the perceptual quality.

Those DNN-based perceptual-oriented quality enhancement methods can be classified into two categories according to the input and output of the enhancement process. The first category is single-frame quality enhancement [6], [7], [8], which is single-frame input and output. The second category is multi-frame assisted quality enhancement [9], [10], which is multi-frame input and single-frame output. Compared with the first category, the second category uses the temporal correlation between frames to improve performance. However, both of these two categories are all frame-by-frame quality enhancement processes, which bring high computational complexity due to the DNN employed to enhance each frame.

Manuscript received 15 June 2023; revised 26 August 2023; accepted 10 September 2023. Date of publication 18 September 2023; date of current version 9 May 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701603, in part by the Natural Science Foundation of China under Grant 62022075 and Grant 61931014, and in part by the Fundamental Research Funds for the Central Universities under Grant WK3490000006. This article was recommended by Associate Editor H. Sun. *(Cunhui Dong and Haichuan Ma contributed equally to this work.) (Corresponding author: Dong Liu.)*

The authors are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: dongcunh@mail.ustc.edu.cn; hcma@mail.ustc.edu.cn; zhuoyuanli@mail.ustc.edu.cn; lil1@ustc.edu.cn; dongeliu@ustc.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3316834.

Digital Object Identifier 10.1109/TCSVT.2023.3316834

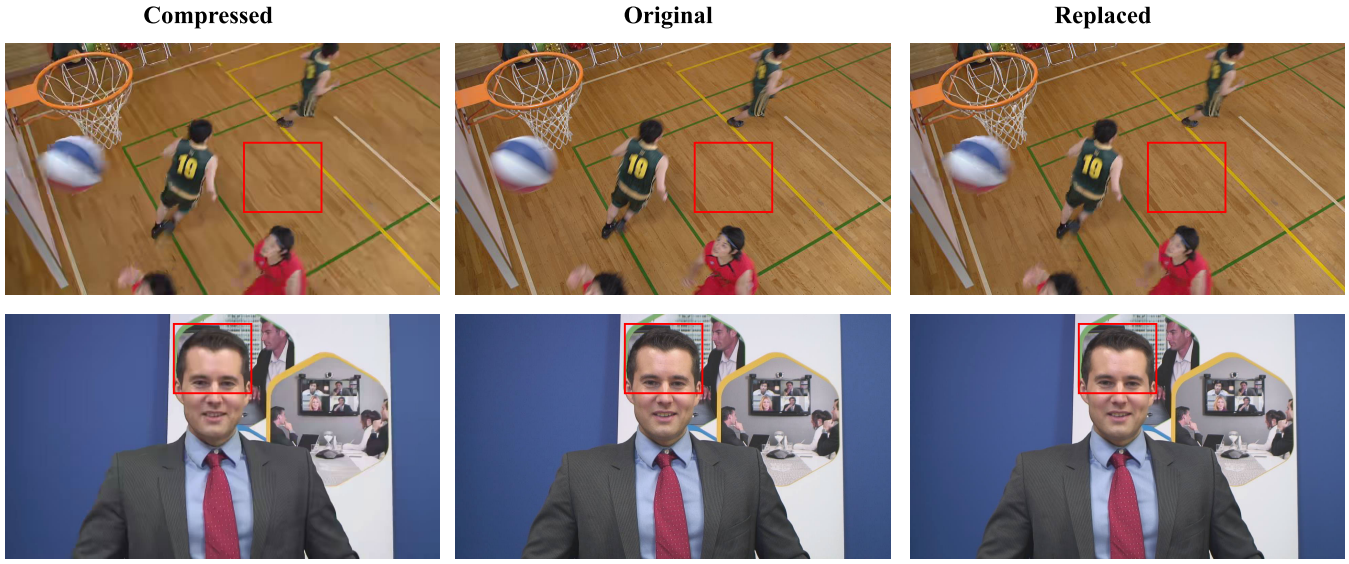


Fig. 1. Visualization of compressed videos, original videos, and replaced videos on the test sequences of JCT-VC [11]. The replaced videos are obtained by replacing the low-frequency frames of compressed videos with those of original videos.

In this paper, we focus on enhancing the perceptual quality of compressed videos with lower complexity. Considering the high temporal correlation between video frames, the temporal low- and high-frequency content indicates the main and variational content of video frames, respectively. Due to the low-frequency content representing the core information of frames, we propose to exclusively enhance the low-frequency content to improve the perceptual quality of frames, which can significantly reduce computational complexity. Fig. 1 shows compressed videos, original videos, and replaced videos which are generated by replacing the compressed low-frequency content with the original low-frequency content. It can be found that the quality of replaced videos is obviously better than that of compressed videos, which represents the significant influence of low-frequency content on perceptual quality. Meanwhile, since only the temporal low-frequency content needs to be enhanced, the amount of computational complexity can be greatly reduced.

Specifically, we propose a temporal frequency-based low-complexity perceptual quality enhancement method. In our method, the temporal wavelet transform (TWT) and temporal wavelet inverse transform (TWIT) are utilized to implement the temporal frequency analysis (TFA) and temporal frequency synthesis (TFS), which are employed to transform the video into temporal frequency frames and transform the temporal frequency frames back into the video, respectively. Different from the spatial wavelet transform which performs wavelet transform directly on the spatial domain, the TWT performs wavelet transform on the temporal domain along the propagation of motion field. Our method composes of three steps. First, a group of compressed frames is fed into the hand-crafted motion estimation (ME) module to acquire the motion field between frames, and then the TWT is employed along the propagation of the motion field to extract temporal high- and low-frequency frames. Second, the temporal low-frequency frames are enhanced by a DNN-based perceptual-oriented quality enhancement network. Finally, TWIT is performed on the temporal high-frequency frames and the enhanced temporal low-frequency frames to generate a group of enhanced compressed frames. Note that one-level TWT transforms two frames into one high-frequency frame and one low-frequency frame. The TWT can further execute multi-scale pyramid decomposition, i.e. n-level TWT is performed on temporal low-frequency frames generated by $(n-1)$-level TWT. Since n-level TWT transforms 2^n compressed frames into one low-frequency frame and 2^n-1 high-frequency frames, our method exclusively enhancing temporal low-frequency frames significantly reduces the computational cost.

In summary, the main contributions of this paper are as follows:

- With the above analysis, we propose to improve perceptual quality by exclusively enhancing low-frequency content, which can exploit the limited computational resource more efficiently. This is the first attempt on addressing the problem of perceptual quality enhancement under computational complexity constraint from the perspective of temporal frequency.
- To achieve this, we propose to implement the TFA and TFS via TWT and TWIT with a hand-crafted ME module. Furthermore, we design a TWT-based low-complexity compressed video perceptual quality enhancement method that can enhance multiple frames simultaneously.
- We conduct extensive experiments to verify the effectiveness of the proposed method. Experimental results show that our method achieves comparable quality enhancement with 13× computational complexity reduction.

The remainder of this paper is organized as follows. Section II gives a brief review of related work. In Section III, we present the temporal wavelet-based method in detail. Section IV analyzes the effect of TFA. In Section V, we show

the performance of our method. In Section VI, we show the additional analysis of ablation studies. Section VII concludes this paper.

II. Related Work

In this section, we review the previous work that relates to our research in two aspects. First, we introduce some methods about the quality enhancement for compressed images/videos. Second, we introduce some studies of wavelet transform-based video compression.

A. Compressed Images/Videos Quality Enhancement

The quantization module is commonly employed with the aim of increasing the compression ratio of images/videos. However, quantization leads to information loss, resulting in the poor quality of compressed images/videos. Therefore, in order to counteract this negative effect, numerous enhancement methods have been proposed to improve the objective and perceptual quality of compressed videos.

First of all, studies focusing on enhancing the objective quality of compressed images/videos are introduced. Dong et al. proposed a four-layer artifact reduction convolutional neural network (AR-CNN) [12] to reduce the compression artifacts in JPEG images. Later, a resource-efficient blind quality enhancement (RBQE) method [13] based on a dynamic CNN architecture was proposed to achieve blind quality enhancement. Moreover, He et al. proposed a deep dual-domain semi-blind network that combined compression quality factor detection and compressed image quality enhancement [14]. For compressed video objective quality enhancement, Dai et al. [1], [15] proposed a variable-filter-size residue-learning CNN-based post-processing approach to replace the traditional loop filter tools (Deblocking [16] and SAO [17]) in HEVC [18], resulting in the higher bit-rate reduction. In [19], a deep CNN-based auto decoder (DCAD) was proposed to reduce the distortion of HEVC compressed videos. Yang *et al*. [2] proposed DS-CNN-I and DS-CNN-B to reduce the intra and inter coding artifacts, respectively. In addition, considering the similarity between consecutive frames, multi-frame quality enhancement (MFQE) [3], [4] was proposed to leverage adjacent frames to assist the enhancement of the current frame and greatly improved objective quality result. Furthermore, Deng et al. introduced a spatiotemporal deformable convolution (STDC) [20] to further aggregate temporal information. These studies on objective quality enhancement aim to minimize pixel-wise loss, such as mean squared error (MSE) and structural similarity index measure (SSIM). However, higher objective quality does not necessarily leads to higher perceptual quality, which has been proved in [5].

To improve the visual experience, various studies have focused on enhancing the perceptual quality of compressed images/videos. Galteri *et al*. [6] presented a fully convolutional residual network trained using a generative adversarial framework for image compression artifact reduction. In [7], taking multiple quality factors into account, Galteri *et al*. proposed an ensemble of GAN driven by a quality predictor, which realistically recovers high-frequency details. For perceptual quality enhancement of compressed videos, CVRGAN [8] firstly introduces GAN loss in network training strategy to enhance the perceptual quality of compressed videos. In [9], Wang et al. demonstrated that spatial high-frequency information had a major impact on the perceptual quality, which performed the wavelet analysis on the spatial domain by DWT (discrete wavelet transform). And they proposed to perform spatial frequency decomposition via wavelet packet to extract the spatial high- and low-frequency content in a frame, and then adopted a reconstruction network with wavelet-dense residual blocks to recover the high-frequency details. Based on MW-GAN, MW-GAN+ [10] introduced an advanced motion alignment network and a 3D discriminator, which achieved a better perceptual enhancement effect. Although these perceptual quality enhancement methods significantly improve the perceptual quality, they can only enhance videos frame-by-frame, which leads to heavy computational complexity.

In addition, different from the previous methods, we adopt temporal wavelet transform in our framework. For the difference of these two transforms (TWT and DWT), DWT is a transform applicable to any discrete sequence. For its application, DWT was usually performed spatially, treating either one row or one column of pixels as a sequence. For example, DWT may decompose a row of pixels into low- and high-frequency components. Therefore, image processing usually adopts 2D DWT, i.e., two DWTs working horizontally and vertically respectively. All these transforms are spatial DWTs. In this paper, we used DWT in the temporal dimension, treating the pixels along a motion thread as a sequence. Thus, it is known as TWT. Such TWTs have been studied for video processing and compression in [21], and [22], for example.

B. Wavelet Transform for Video Compression

Wavelet transform is a powerful signal processing tool, which has been extensively applied in numerous fields. In video coding, due to its frequency decomposition characteristics and ability to avoid block artifacts, various wavelet-based video coding methods have been proposed and intensively studied. Hsiang *et al*. proposed a wavelet-based motion compensated embedded zero block scalable video coder (MC-EZBC) in [23], which was designed based on the embedded image coding scheme EZBC and utilized an invertible motion-compensated 3-D wavelet subband filter bank for video analysis/synthesis. Based on MC-EZBC, Wu *et al*. proposed the well-known enhanced motion-compensated embedded zero block coding (ENH-MC-EZBC) [24], which significantly outperformed MC-EZBC. In [25], Chen et al. proposed a content adaptive Lagrange multiplier selection algorithm for the wavelet-based video coding, which enabled the more satisfactory video quality with negligible additional computational complexity. Inspired by [26], and [27], Dong et al. [22] proposed a learnable wavelet inverse transform for scalable video coding, significantly improving compression performance.

In addition to the wavelet-based video coding of the MC-EZBC series, various other wavelet-based video coding

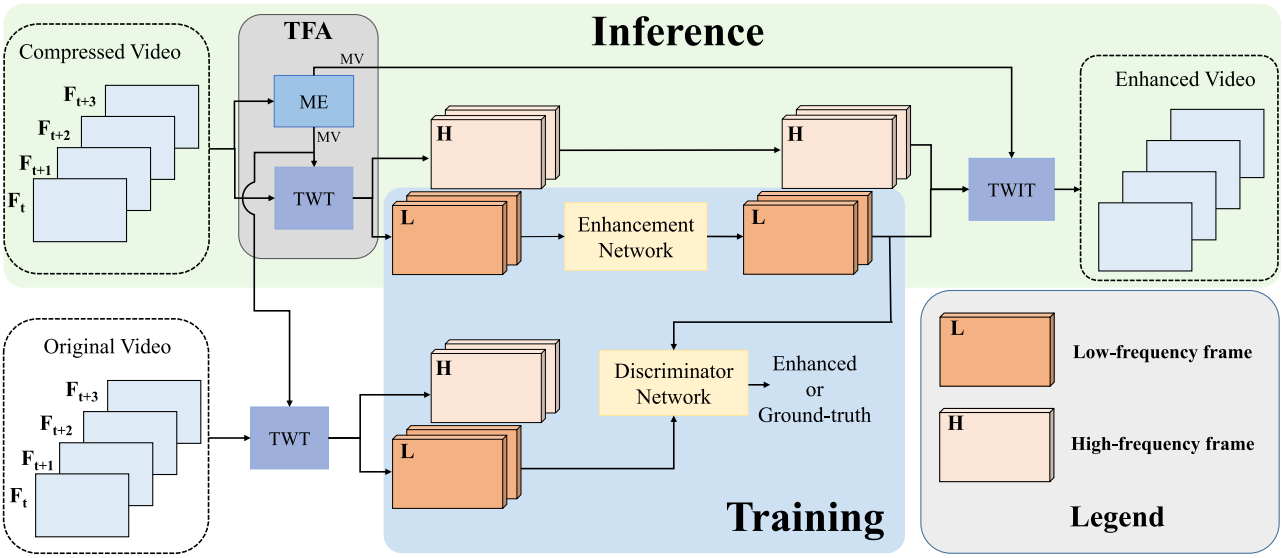


Fig. 2. The one-level temporal wavelet-based low-complexity perceptual quality enhancement framework from the perspectives of inference and training. The TFA indicates temporal frequency analysis, it includes two modules: motion estimation (ME) and temporal wavelet transform (TWT), where the ME module is used to extract the motion vector (MV) between video or low/high-frequency frames, and the TWT module is used to generate the temporal high- and low-frequency frames by performing the wavelet transform on video frames along the propagation of MV. During inference (green box), the compressed video is fed into the TFA to generate temporal high- and low-frequency frames, and low-frequency frames are enhanced by the enhancement network. Then the enhanced temporal low-frequency frames and high-frequency frames are fed into the temporal wavelet inverse transform (TWIT) to generate the enhanced videos by performing the inverse wavelet transform along the propagation of MV. During training (blue box), TWT is performed on the original video along the MV (extracted from the compressed video) to generate the training labels. The original and compressed low-frequency frames form a training pair to train the network in an adversarial manner.

methods have been proposed. In [28], Onthriar *et al*. proposed a wavelet-based Dirac video codec which employed the motion compensation and 2-D wavelet transform. Based on Dirac video codec, Dam *et al*. investigated a nonlinear quantization method [29], resulting in significant improvements in compression performance. In [30], Bystrov *et al*. implemented a multichannel wavelet video codec based on the Schrodinger codec, which was an optimized version of the Dirac video codec. Owing to the more compact representation of the signal energy along the frequency subbands, the multichannel wavelet significantly improved the compression performance. In [31], Jin *et al*. utilized the DWT to obtain the hierarchy of multi-frequency components at different spatial resolutions, and further efficiently extracted the structural and detailed information in temporal context to boost the motion vector and residue compression. In [32], Meyer *et al*. proposed an end-to-end trainable wavelet video coder based on motion-compensated temporal filtering (MCTF), which considered multiple temporal decomposition levels in MCTF during training and adapted to different motion strengths during inference.

The aforementioned studies predominantly utilize wavelet transform as the transform module for video coding. For quality enhancement of compressed videos, Wang *et al*. [9], [10] proposed to employ the spatial wavelet packet to decompose a frame into high- and low-frequency content, and then recover the high-frequency details to enhance the perceptual quality. These methods verify the potential of the frequency-based thinking in quality enhancement. However, no studies further investigated the quality enhancement of compressed videos in the temporal frequency-domain, and further excavate the frequency analysis properties of the temporal-domain.

III. Proposed Method

In this section, we introduce the proposed temporal wavelet-based low-complexity perceptual quality enhancement method. The overview of the method is presented first, and then the details of the method are introduced.

A. Overview

Fig. 2 shows the one-level TWT-based perceptual-oriented quality enhancement method from the perspectives of both inference and training.

During inference (illustrated by green box in Fig 2), it mainly includes three steps. First, the compressed video is fed into the TFA module to perform temporal analysis. Specifically, the ME is employed to estimate the motion vector (MV) of the compressed video, and then the TWT module is utilized to extract temporal high-frequency frames H and temporal low-frequency frames L by performing the wavelet transform on compressed videos along the propagation of MV. Second, the temporal low-frequency frames L are enhanced by using the perceptual-oriented quality enhancement network. Third, the temporal high-frequency frames H, the enhanced temporal low-frequency frames L, and the MV are fed into the TWIT to perform the TFS. By the TWIT, the temporal wavelet frames are converted to the enhanced video. Note that only one-level TFA is shown in Fig. 2, the multi-level pyramid TFA can be performed in the actual process, i.e. the next level TFA performs ME and TWT on temporal low-frequency frames generated by the current level TFA.

During training (illustrated by blue box in Fig 2), in order to get the training labels of the enhancement network, TWT is performed on the original video along the propagation

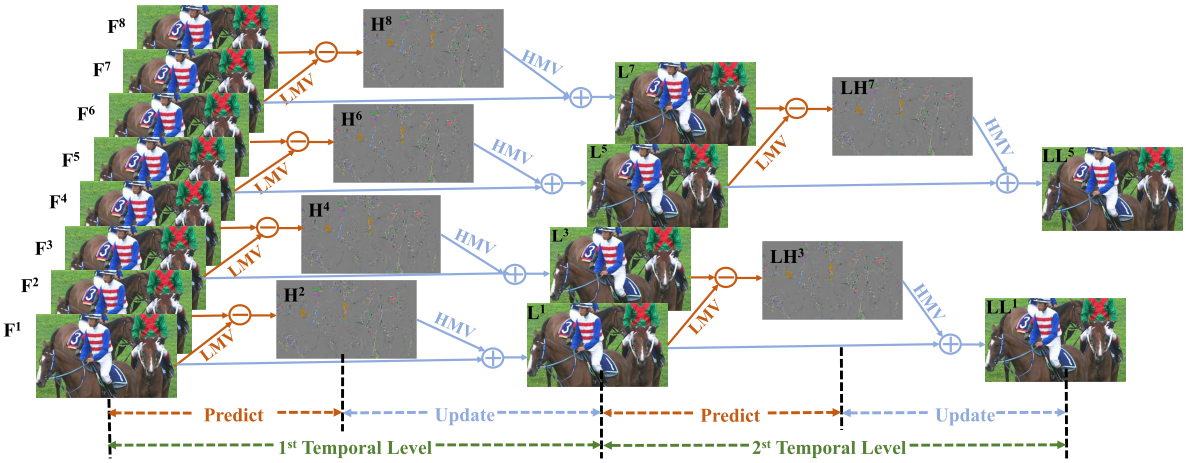


Fig. 3. The pipeline of 2-level temporal forward Haar wavelet transform. LMV represents the MV between video/low-frequency frames, while HMV represents the MV between high-frequency frames and video/low-frequency frames. The branches of LMV and HMV represent motion compensation along the MV of LMV and HMV, respectively.

of MV from the compressed video to obtain the original low-frequency frames. The original low-frequency frames and compressed low-frequency frames form a training pair to train the enhancement network in an adversarial manner.

As shown in Fig. 2, our framework mainly comprises three modules: TWT, ME, and the enhancement network. The TWT module is employed for frequency analysis by performing the wavelet transform along the propagation of MV. It will be introduced in Section III-B. The MV is generated through the hand-crafted ME module which will be introduced in Section III-C. The enhancement network is utilized to enhance the low-frequency frames, the details regarding the network structure, training strategy, and other relevant information will be introduced in Section III-D.

B. Temporal Wavelet Transform

In this subsection, we introduce the TWT and TWIT modules. There are three types of TWT: temporal forward Haar wavelet, temporal backward Haar wavelet, and temporal CDF 5/3 wavelet. Among these types, a suitable one is selected as required. First, we introduce a detailed account of the TWT using the temporal forward Haar wavelet as an example. Second, we introduce the remaining two TWTs, which are similar to the temporal forward Haar wavelet. In addition, we present how to select the type of TWT. Finally, the TWIT is introduced.

1) Temporal Forward Haar Wavelet Transform: The temporal forward Haar wavelet transform is designed based on the lifting wavelet. Compared with the first generation wavelet, the lifting wavelet offers a faster and in-place implementation. Fig. 3 shows the 2-level lifting wavelet-based temporal forward Haar wavelet transform. The transform comprises three fundamental stages: split, prediction, and update. For split, since the wavelet transform is performed in the temporal dimension, the split divides frames into update frames $\{F_1, F_3, F_5, F_7\}$ and prediction frames $\{F_2, F_4, F_6, F_8\}$ according to the parity of time. For prediction and update, different from the spatial wavelet transform that directly filters on images, the TWT needs to consider the propagation of motion field between different video frames. Therefore, for the prediction of TWT, pixels of the prediction frame are employed to subtract pixels offset by MV within the update frame. Similarly, for the update, the pixels of the update frame are employed to subtract the pixels offset by MV within the high-frequency frame. The specific formulas of prediction and update are as follows:

$$\begin{aligned} H^{2t}[m,n] &= F^{2t}[m,n] - F^{2t-1}[m - LMV_x^{2t-1}[m,n], n \\ &\quad - LMV_y^{2t-1}[m,n]] \\ L^{2t-1}[m,n] &= F^{2t-1}[m,n] + \frac{1}{2} \times H^{2t}[m - HMV_x^{2t}[m,n], n \\ &\quad - HMV_y^{2t}[m,n]] \end{aligned} \tag{1}$$

where F^{2t} and F^{2t-1} represent the frame at time $2t$ and $2t-1$, respectively. H^{2t} represents the high-frequency frame obtained by prediction, and L^{2t-1} represents the low-frequency frame obtained by update. m and n represent the vertical and horizontal coordinates of pixels. LMV^{2t-1} represents the MV between the current frame and the frame F^{2t-1}, and HMV^{2t} represents the MV between the current frame and the high-frequency frame H^{2t}, which will be introduced in Section III-C. In order to obtain multi-level pyramid frequency analysis, multi-level TWT is employed, which performs TWT on the previous low-frequency frames. For example, Fig. 3 shows the second-level TWT which performs on the temporal low-frequency frames $\{L_1, L_3, L_5, L_7\}$.

2) Other Types of TWT and the Selection of Different TWTs: To reduce computational complexity, our method exclusively enhances temporal low-frequency frames. It is expected that information be concentrated in low-frequency frames as much as possible, leaving less high-frequency information in high-frequency frames, so the quality of all video frames can be better enhanced. In this subsection, to concentrate more information in low-frequency frames, we additionally employ the temporal backward Harr wavelet transform and the temporal CDF 5/3 wavelet transform. They are similar to the temporal forward Haar wavelet transform, which are also designed based on the lifting wavelet and require three steps of split, prediction, and update. The differences among the three TWTs

are inputs and formulas for prediction and update, which will be introduced in detail below. Then we describe how to select the type of TWT.

The temporal forward Haar wavelet employs $2t$ and $2t-1$ frames for prediction and update, as shown in Eq. 1. However, some content in the current frame may be occluded in the previous frame, but not occluded in the following frame. Therefore, the temporal backward Haar wavelet which employs $2t$ and $2t+1$ frames is proposed. Its formulas for prediction and update are:

$$\begin{aligned} H^{2t}[m,n] &= F^{2t}[m,n] - F^{2t+1}[m - LMV_x^{2t+1}[m,n], n \\ &\quad - LMV_y^{2t+1}[m,n]] \qquad (2) \\ L^{2t+1}[m,n] &= F^{2t+1}[m,n] + \frac{1}{2} \times H^{2t}[m - HMV_x^{2t}[m,n], n \\ &\quad - HMV_y^{2t}[m,n]] \qquad (3) \end{aligned}$$

Bi-directional prediction is an important inter-frame prediction technique in video coding. Compared with unidirectional prediction which exclusively utilizes the previous or the following frame, bidirectional prediction utilizes both the previous and the following frames, so bidirectional prediction can significantly improve the prediction accuracy and reduce the residual. Based on the idea of bidirectional prediction, we introduce the temporal CDF 5/3 wavelet to reduce the information in high-frequency frames. The temporal CDF 5/3 wavelet employs both the previous frame and the following frame. The formulas for prediction and update are as follows:

$$\begin{aligned} H^{2t}[m,n] &= F^{2t}[m,n] - \frac{1}{2} \times F^{2t-1}[m - LMV_x^{2t-1}[m,n], n \\ &\quad - LMV_y^{2t-1}[m,n]] \\ &\quad - \frac{1}{2} \times F^{2t+1}[m - LMV_x^{2t+1}[m,n], n \\ &\quad - LMV_y^{2t+1}[m,n]] \\ L^{2t-1}[m,n] &= F^{2t-1}[m,n] + \frac{1}{4} \times H^{2t}[m - HMV_x^{2t}[m,n], n \\ &\quad - HMV_y^{2t}[m,n]] \\ &\quad + \frac{1}{4} \times H^{2t-2}[m - HMV_x^{2t-2}[m,n], n \\ &\quad - HMV_y^{2t-2}[m,n]] \qquad (4) \end{aligned}$$

There exist three distinct TWTs, but only one among them is selected for transform. The selection of different TWTs is pixel-wise and the selection process consists of three steps. First, three high-frequency frames are generated by utilizing three TWTs. Second, each pixel in the prediction frame selects the prediction type that produces the minimum absolute value in the high-frequency frame. Third, the update type of each pixel in the update frame is selected according to whether the current pixel participates in the prediction of the previous frame or the following frame. If a pixel participates in both the prediction of the previous frame and the following frame, the temporal CDF 5/3 wavelet update is selected. However, if the pixel only participates in the prediction of the previous frame, then the temporal backward Haar wavelet update is selected, and vice versa.

3) TWIT: TWIT is utilized for converting temporal wavelet frames back into reconstructed videos. TWIT is designed as a perfect inversion of the TWT, thus TWIT has three types corresponding to the three types of TWT. The type of TWIT is consistent with that of TWT and the formula of TWIT can be derived from the formula of TWT. For example, if TWT uses the temporal forward Haar wavelet, the TWIT also uses the temporal forward Haar wavelet, and the specific formula is derived from (Eq. 1), the details as follows:

$$\begin{aligned} F^{2t-1}[m,n] &= L^{2t-1}[m,n] - \frac{1}{2} \times H^{2t}[m - LMV_x^{2t}[m,n], n \\ &\quad - LMV_y^{2t}[m,n]] \\ F^{2t}[m,n] &= H^{2t}[m,n] + F^{2t-1}[m - HMV_x^{2t-1}[m,n], n \\ &\quad - HMV_y^{2t-1}[m,n]] \qquad (5) \end{aligned}$$

Similarly, the formulas of the other two TWITs can be derived by (Eq. 2) and (Eq. 4), respectively.

C. Motion Estimation

ME is an important tool for motion alignment between frames. As such, diverse ME methods have been developed, including deep learning-based ME methods [33], [34], [35] and hand-crafted ME methods [36], [37]. Considering the high computational complexity of deep learning-based methods, in our method, the hand-crafted ME method with lower computational complexity is adopted. A mature hand-crafted method is the ME module [38], [39] in traditional video coding frameworks [18], [40]. However, the ME in video coding frameworks commonly considers the bit-rate of coding MV, which not only increases the computational complexity but also leads to a poor alignment effect. Instead of directly using the existing ME module, we implement a ME module to fit our quality enhancement scheme.

In our method, the TWT requires two MVs: LMV and HMV. The LMV represents the motion vector between video/low-frequency frames, while the HMV represents the motion vector between the high-frequency frame and the video/low-frequency frame. For the LMV, we implement the hand-crafted ME module based on traditional block matching which utilizes the absolute difference sum (SAD) of pixels between the current block and the reference block as the constraint condition. In addition, in order to reduce the error of ME, we incorporate MV smoothness constraint condition. For the HMV, it cannot be directly searched through the SAD-based ME. In our method, the HMV is generated by reversing the LMV. The specific calculation methods of LMV and HMV are introduced in detail below.

1) LMV: The three-step search algorithm is proposed as the search method, which offers the advantages of high accuracy and efficiency. In light of the greater search ratio afforded by hexagons, it is utilized as the search template. Considering the sub-pixel motion and the complexity of interpolation, we opt for the bilinear interpolation method. In addition, to start from a better initial MV, we utilize the MVP technique which employs the MVs of neighboring blocks to determine the starting point of the search.

Fig. 4. This figure illustrates the MV generated based only on SAD constraint, and the MV generated based on the integration of SAD and MV smoothness constraints. Sole reliance on SAD may result in inaccurate results. However, by imposing constraints on the smoothness of the MV, the correct MV can be determined.

The SAD-based ME exclusively considers pixel errors, which may cause matching inaccuracy on small blocks, as shown in Fig. 4. The red block in the current frame ought to match the red block in the reference frame, but it matches the green block because of the slightly smaller SAD. To solve this problem, we incorporate MV smoothness constraints that promote the similarity between the searched MV and its neighboring MVs, thereby reducing the probability of mismatch. The smoothness constraints are achieved in two ways. The first way is the variable block size technique. Specifically, if the SAD between the searched optimal reference block and the current block exceeds the predetermined threshold ϵ, the current block is divided into a quadtree, then ME is executed on each sub-block. However, if the SAD is less than the threshold ϵ, the current block directly employs the searched MV. The second way is that when calculating the cost of ME, not only the SAD between the matching block and the current block is considered, but also the displacement between the current MV and its adjacent MVs is taken into count. The specific cost formula is as follows:

$$Cost = SAD(curBlock, refBlock) + \lambda \times MVD \tag{6}$$

where MVD represents the absolute difference between the current MV and its adjacent MVs, and λ is utilized to control the degree of smoothness.

2) HMV: The HMV employed by the update in TWT represents the MV between the high-frequency frame and the video/low-frequency frame. Due to the coefficients of the high-frequency frame are high-frequency coefficients, HMV cannot be generated by the pixel value matching-based ME, as used by the LMV. Drawing on the concept that the backward optical flow is the inversion of the forward optical flow, we derive the HMV from its corresponding LMV, which is temporal aligned with the HMV. For example, in Fig. 3, the HMV between H^2 and F^1 corresponds to the LMV between F^2 and F^1. The specific derive formula is as follows:

$$\begin{aligned} m_u &= m + \lfloor LMV_x[n,m] \rceil \\ n_u &= n + \lfloor LMV_y[n,m] \rceil \\ HMV_x[n_u,m_u] &= -LMV_x[n,m] \\ HMV_y[n_u,m_u] &= -LMV_y[n,m] \end{aligned} \tag{7}$$

where the $\lfloor \cdot \rceil$ is round function. m_u and n_u beyond the frame boundary are directly discarded. In addition, some positions in the HMV may not be assigned, which indicates that the correlation between frames is weak. So the update operation exclusively is executed for the assigned positions in HMV.

D. Network Structure and Loss Function

1) Network Structure: Fig. 5 illustrates the architectures of the enhancement network and the discriminator network. The enhancement network employs a plain CNN with multiple residual blocks to facilitate effective training. The discriminator network consists of three convolutional layers and one max pooling operation. By the downsample of max pooling, the discriminator achieves a larger receptive field.

2) Loss Function: To obtain the training label of the network, we conduct TWT on the original video to generate the original temporal low-frequency frame L. The pair of L and the compressed temporal low-frequency frame $\hat{L}$ are employed as training data for the enhancement network. The loss function of the enhancement network consists of three components: the adversarial loss $L_{G-Adv}(\theta_G)$, the feature domain loss $L_{Fea}(\theta_G)$, and the wavelet domain loss $L_{Wav}(\theta_G)$. In addition, the loss function of the discriminator network is the adversarial loss $L_{D-Adv}(\theta_D)$. Here, θ_G and θ_D refer to the parameters of the enhancement and discriminator networks, respectively. Below are detailed explanations of these loss functions.

a) Adversarial loss: Studies [41], [42] have demonstrated that the adversarial loss optimizes the distance between the distribution of generated images and the training set. Moreover, it has been demonstrated in [5] that the distance between distributions is associated with the perceptual quality, and a smaller distance leads to a better perceptual quality. Numerous studies [43], [44], [45] also have verified that incorporating adversarial loss can improve the perceptual quality. In our method, we employ the generative adversarial training technique proposed in the paper [46], which is simple for training and performs effectively. Specifically, the $L_{G-Adv}(\theta_G)$ and $L_{D-Adv}(\theta_D)$ are as follows,

$$\begin{aligned} L_{G-Adv}(\theta_G) = &\frac{1}{2} \times E[\left| D(G(\hat{L})) - D(L) \right|] \\ &+ \frac{1}{2} \times E[D(L) - D(G(\hat{L}))] \end{aligned} \tag{8}$$

$$L_{D-Adv}(\theta_D) = -E[D(L) - D(G(\hat{L}))] \tag{9}$$

b) Feature loss: To further improve the perceptual quality, similar to the previous work [44], we calculate the difference between the deep features of the enhanced frame $G(\hat{L})$ and the original label L. The deep feature map of the pre-trained VGG-19 [47] network before the activation layer is employed to calculate the feature distortion, specifically as shown in the following formula,

$$L_{Fea}(\theta_G) = \left\| f_{VGG}(L) - f_{VGG}(G(\hat{L})) \right\|_1 \tag{10}$$

where $f_{VGG}()$ represents the feature maps from VGG-19.

c) Wavelet loss: Considering the training of GAN is not stable, in addition, we employ the wavelet domain loss to stabilize the training process. Specifically, we utilize the *Charbonnier* distance [48] between the original label L and

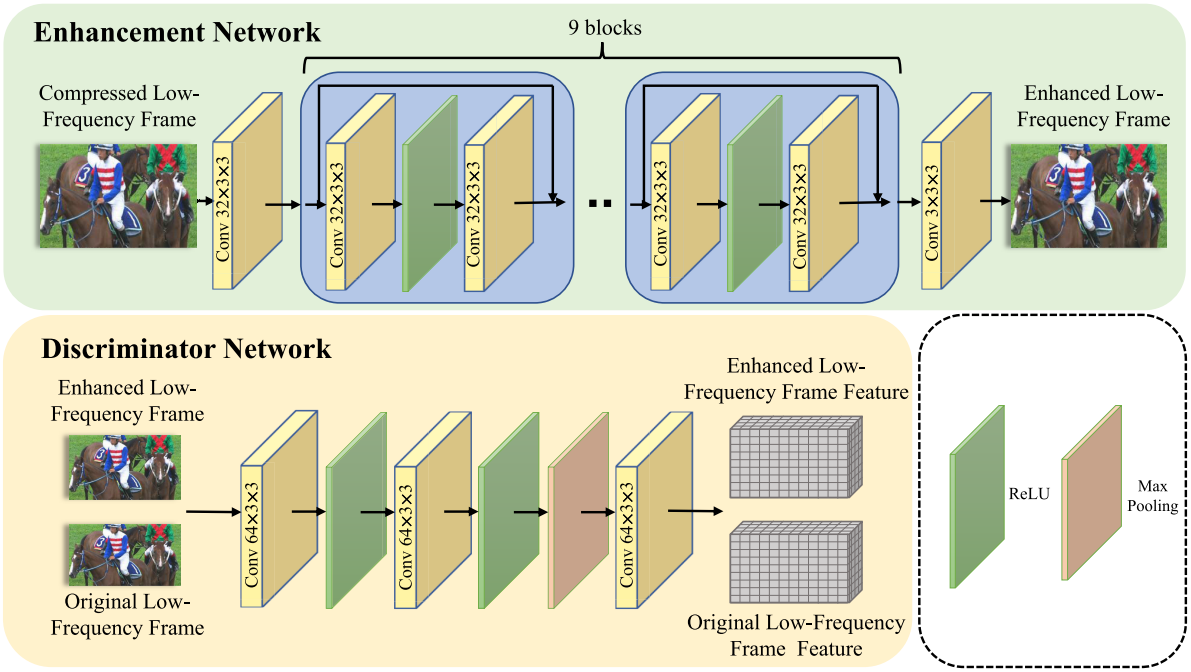


Fig. 5. Architectures of the enhancement network and the discriminator network.

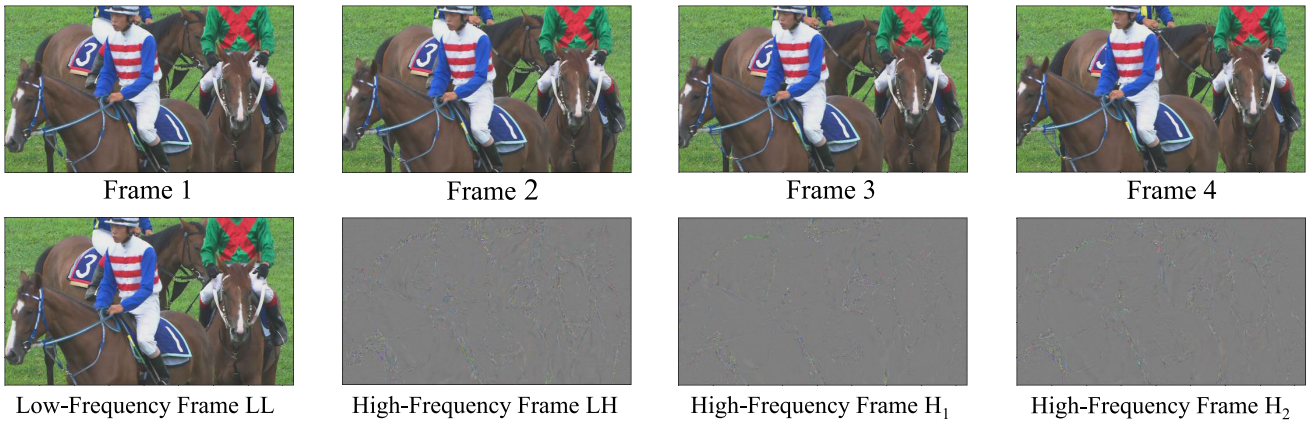


Fig. 6. The temporal low-frequency frame and three temporal high-frequency frames obtained by the 2-level TWT on four frames.

the enhanced frame $G(\hat{L})$ as the wavelet domain loss. The specific formula is as follows,

$$L_{Wav}(\theta_G) = \sqrt{\left\| L - G(\hat{L}) \right\|_2 + \epsilon^2} \tag{11}$$

where $\|.\|_2$ represents the 2-norm, ϵ is a very small hyperparameter.

d) Overall loss: The overall loss function of the enhancement network is a weighted combination of the adversarial loss, the feature loss, and the wavelet loss, as shown below,

$$\min_{\theta_G} : \alpha \times L_{Fea}(\theta_G) + \beta \times L_{Wav}(\theta_G) + \gamma \times L_{G-Adv}(\theta_G) \tag{12}$$

where α, β, and γ are hyper-parameters controlling the weight of the three losses. The loss function of the discriminator network is (Eq. 9). The θ_G and θ_D are alternately optimized by gradient descent for adversarial training.

IV. Analyzing the Effect of TFA

In this section, we focus on analyzing the effect of TFA. To verify the conjecture that the temporal low-frequency content is the main content of the video and the temporal high-frequency content is the variation between frames, we perform 2-level TWT on four video frames, yielding a temporal low-frequency frame LL and three temporal high-frequency frames $\{LH, H_1, H_2\}$. The results are illustrated in Fig. 6. It can be found that the main content of the video concentrates on the low-frequency frame, while the high-frequency frames are some residual information.

We further design a replacement experiment to quantitatively verify that exclusively enhancing low-frequency frames can improve the perceptual quality of the overall video. The replacement experiment has three steps. First, we perform 4-level TWT on compressed videos to obtain low-frequency frames LF_{com} and high-frequency frames HF_{com}. Second, similarly, original low-frequency frames LF_{org} and

TABLE I
THE AVERAGE PI [49] RESULT OF COMPRESSED VIDEOS, REPLACED VIDEOS AND ORIGINAL VIDEOS ON THE TEST SET OF JCT-VC [11]. REPLACED VIDEOS ARE GENERATED BY REPLACING LOW-FREQUENCY FRAMES OF COMPRESSED VIDEOS WITH THOSE OF ORIGINAL VIDEOS

Method	Compressed	Replaced	Original
PI	4.68	3.37	3.32

high-frequency frames HF_{org} are obtained by conducting 4-level TWT on original videos via the MV of compressed videos. Finally, compressed low-frequency frames LF_{com} are replaced with original low-frequency frames LF_{org}, that is, reconstructed videos are generated by performing TWIT on LF_{org} and HF_{com}. The experiment is conducted on the test set of JCT-VC [11], and compressed videos are obtained by HM-16.5 [18] with Low Delay (LD) configuration at QP = 37. The average perceptual index (PI) [49] index of compressed videos, replaced videos, and original videos are shown in Table I. Compared with the PI of compressed videos, the PI of the replaced videos is significantly lower, which is almost the same as that of original videos. This shows that the low-frequency frames obtained by the TWT are the main content of videos, and exclusively enhancing the low-frequency frames can improve the perceptual quality of videos. In Fig. 1, we visualize the frames of compressed videos, replaced videos, and original videos. As shown in the figure, replaced videos are more fidelity to original videos, and significantly better than the compressed videos.

V. EXPERIMENTS

A. Settings

1) Datasets: In order to include videos of different contents and resolutions, we construct a composite training dataset that comprises three parts: MFQE [4], TVD [50], BVI-DVC [51]. All sequences are compressed by HM-16.5 [18] under LD configuration with QPs = 37 and 42, respectively.

To evaluate the performance of our method, we test our method and other quality enhancement methods on the standard test sequences of JCT-VC [52]. All test sequences are compressed by HM-16.5 under LD configuration with QPs = 37 and 42, respectively.

2) Implementation Details: The followings are some hyper-parameters and training strategies in our experiments. We employ the 4-level TWT which transforms sixteen frames into one low-frequency frame and fifteen high-frequency frames. The training process of the enhancement network is divided into two stages. In the first stage, the training loss function only includes wavelet domain loss, that is, the α, β, γ in the (12) are set to 1, 0, 0, respectively. Adam optimizer [53] with $lr = 5 \times 10^{-5}, \beta_1 = 0.9$ and $\beta_2 = 0.999$ is utilized to train the enhancement network. In the second stage, the pre-trained model in the first stage is loaded as an initialization. The training loss function contains the wavelet domain loss, the feature domain loss, and their adversarial loss, and the weights α, β, γ are set to $1 \times 10^{-4}, 1 \times 10^{-2}, 1 \times 10^{-2}$, respectively. RMSprop optimizer [54] with $lr = 5 \times 10^{-5}, \alpha = 0.9$ and $\epsilon = 10^{-8}$ is utilized to train the enhancement network and discriminator network. The enhancement network and discriminator network are optimized in an alternating manner. The discriminator network is optimized in five rounds first, then the enhancement network is optimized in one round.

B. Quantitative Comparison

In this subsection, we conduct a quantitative comparison of our proposed method with other quality enhancement methods. Similar to prior studies [9], [10], we employ the learned perceptual image patch similarity (LPIPS) [55] and PI [49] as the evaluation criteria. These metrics are commonly employed for perceptual quality evaluation. We compare our proposed method with MFQE 2.0 [4], CVRGAN [8], MW-GAN [9], and MW-GAN+ [10]. Among them, MFQE2.0 [4] is an excellent objective quality enhancement method. CVRGAN [8] firstly employs the adversarial loss in perceptual quality enhancement on compressed videos and demonstrates notable improvement in perceptual quality for compressed videos. MW-GAN introduces multi-level wavelet-based decomposition, which helps to significantly improve the perceptual quality. Furthermore, MW-GAN+ is an enhanced version of MW-GAN, which has incorporated many significant technologies and exhibited superior performance over MW-GAN.

Table II presents the ΔLPIPS and ΔPI metrics calculated between enhanced videos and compressed videos on the test set of JCT-VC [11]. Note that the decreased LPIPS and PI indicate better perceptual quality. As shown in table II, our method outperforms all other methods in terms of LPIPS and PI. Specifically, the average ΔLPIPS and ΔPI of our method at QP = 42 are −0.065 and −1.54, which are 18.2% and 13.2% better than that of MW-GAN+. At QP = 37, the average ΔLPIPS and ΔPI of our method are −0.064 and −1.38, which are 20.7% and 20.5% better than that of MW-GAN+. In contrast, the MFQE which mainly focuses on objective quality enhancement has positive ΔLPIPS and ΔPI values, indicating the degradation of perceptual quality.

C. Subjective Comparison

In this subsection, we focus on the subjective comparison of our proposed method with other methods. Fig. 7 visualizes the subjective comparison of different method, including compressed videos, MW-GAN, MW-GAN+, our method, and raw videos. The sequences of RaceHorses, PeopleOnStreet, BasketballDrill, and Traffic from JCT-VC [11] are shown in Fig. 7. It can be found that the videos processed by our proposed method have sharper edges and better perceptual qualities.

To further evaluate the subjective quality of our method, we conduct the mean opinion score (MOS) test. Fifteen subjects participate in the MOS test, and the dual incentive approach is employed. Each subject rates an integral score (from 1 to 5) for each video and a higher score indicates better perceptual quality. The test set is the sequences of JCT-VC [11] and is compressed by HM-16.5 with QP = 37. A "center crop" is applied to videos with resolutions exceeding

TABLE II
OVERALL ΔLPIPS AND ΔPI BETWEEN COMPRESSED VIDEOS AND ENHANCED VIDEOS ON THE TEST SET OF JCT-VC [11]

QP	Video sequence		MFQE		CVRGAN		MW-GAN		MW-GAN+		Ours	
			ΔLPIPS	ΔPI	ΔLPIPS	ΔPI	ΔLPIPS	ΔPI	ΔLPIPS	ΔPI	ΔLPIPS	ΔPI
42	A	Traffic	0.055	0.566	-0.04	-0.452	-0.034	-0.913	-0.056	-1.399	-0.055	-1.422
		PeopleOnStreet	0.038	1.132	-0.031	-0.2	-0.003	-0.648	-0.016	-0.69	-0.014	-1.408
	B	Kimono	0.073	0.438	-0.03	-1.346	-0.018	-2.564	-0.076	-1.45	-0.029	-1.343
		ParkScene	0.103	0.844	-0.053	-0.895	-0.017	-0.222	-0.097	-0.699	-0.119	-1.9
		Cactus	0.051	0.531	-0.054	-0.716	-0.052	-1.021	-0.092	-1.534	-0.107	-1.678
		BQTerrace	0.075	0.359	-0.022	-0.26	-0.025	-0.53	-0.092	-0.798	-0.057	-0.785
		BasketballDrive	0.077	0.94	-0.044	-0.497	-0.041	-0.816	-0.083	-1.015	-0.102	-1.228
	C	RaceHorses	0.071	1.218	-0.023	-0.548	-0.016	-1.229	-0.034	-1.514	-0.072	-2.16
		BQMall	0.041	1.214	0	-0.312	-0.015	-0.887	-0.046	-1.157	-0.072	-1.26
		PartyScene	0.058	0.808	0.007	-0.379	-0.06	-1.549	-0.05	-1.762	-0.085	-0.585
		BasketballDrill	0.056	0.753	-0.033	-0.811	-0.029	-0.775	-0.04	-0.894	-0.056	-1.349
	D	RaceHorses	0.066	1.124	-0.016	-0.416	-0.008	-1.183	-0.053	-1.317	-0.054	-1.893
		BQSquare	0.088	0.867	0.015	-0.513	-0.003	-0.837	-0.016	-1.316	-0.044	-1.028
		BlowingBubbles	0.053	0.656	-0.006	-0.596	-0.043	-1.485	-0.087	-2.054	-0.094	-2.202
		BasketballPass	0.051	1.01	-0.006	-1.183	-0.01	-1.349	-0.023	-1.999	-0.060	-2.591
	E	FourPeople	0.032	0.271	-0.002	-0.546	-0.022	-1.565	-0.037	-2.014	-0.038	-1.539
		Johnny	0.045	0.43	-0.029	-0.816	-0.063	-1.431	-0.064	-1.263	-0.062	-1.416
		KristenAndSara	0.045	0.346	-0.013	-0.814	-0.036	-1.57	-0.034	-1.608	-0.044	-1.881
	Average		**0.06**	**0.75**	**-0.021**	**-0.626**	**-0.025**	**-1.143**	**-0.055**	**-1.36**	**-0.065**	**-1.54**
37	**Average**		**0.024**	**0.614**	**-0.021**	**-0.513**	**-0.046**	**-0.993**	**-0.053**	**-1.145**	**-0.064**	**-1.38**

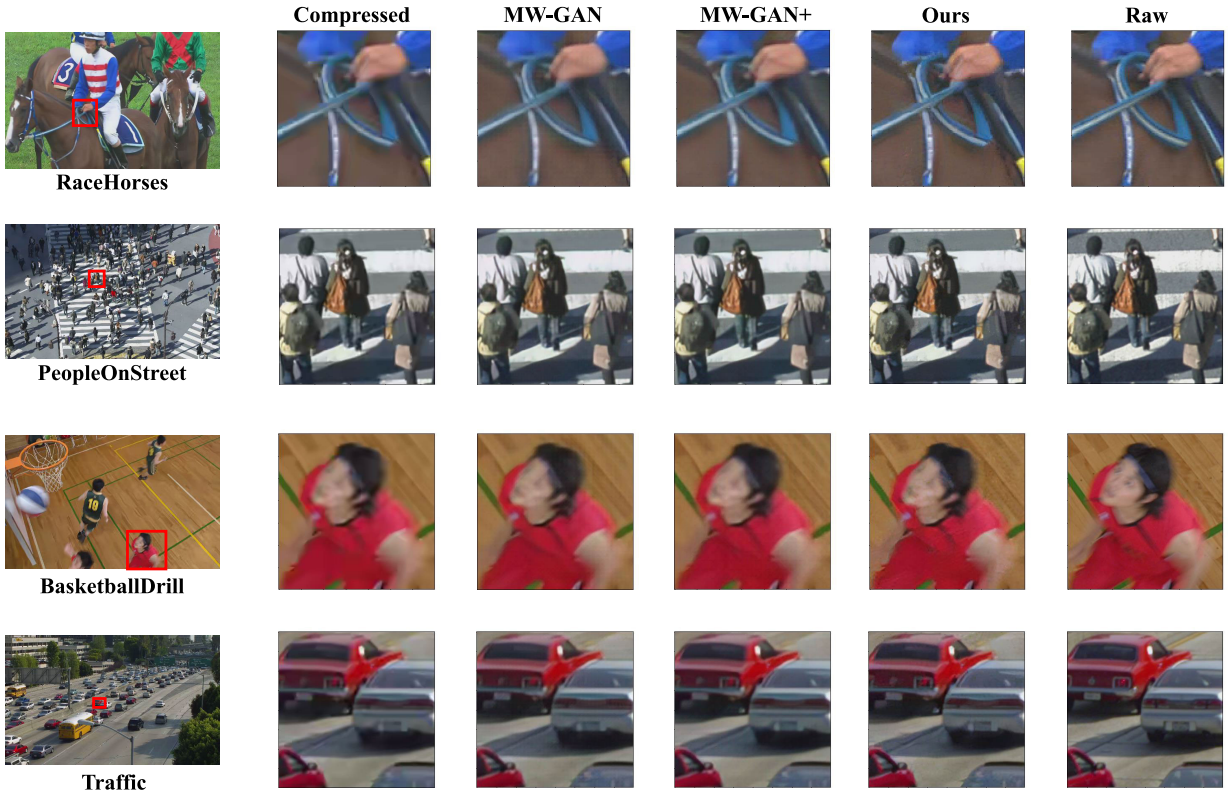


Fig. 7. Visualization comparison of compressed videos, videos enhanced by MW-GAN, videos enhanced by ME-GAN+, videos enhanced by our method, as well as raw videos on the test sequences of JCT-VC [11] (Zoom in for best view).

1080p (e.g., Class A). Subjects are instructed to give scores for the compressed videos, enhanced videos by MW-GAN+ [10], and enhanced videos by our method. MOS results are presented in Table III. Compared with compressed videos,

TABLE III
THE COMPREHENSIVE RESULTS (MOS, ΔLPIPS, FLOPS) OF COMPRESSED VIDEOS, VIDEOS ENHANCED BY MW-GAN+, AND VIDEOS ENHANCED BY OUR METHOD ON THE TEST SET OF JCT-VC [11]

Metric	Compressed	MW-GAN+	Ours
MOS	2.95	3.29	3.26
ΔLPIPS	-	-0.053	-0.064
ΔPI	-	-1.145	-1.380
FLOPs	-	4995.98G	109G

videos enhanced by our method achieve a significant improvement in perceptual quality. In addition, compared with the state-of-the-art method (MG-GAN+), the computational complexity of our method is much less than the state-of-the-art method, while achieving better performance on LPIPS and PI and comparable performance on MOS.

D. Computational Complexity

In this subsection, we compare the computational complexity (floating point operations, FLOPs) of our method with other perceptual quality enhancement methods. The computation burden of our method primarily comprises three components: ME, TWT & TWIT, and the enhancement of low-frequency frames.

1) ME: It consists of three stages: interpolation, block-based ME for LMV, and HMV generation. For the interpolation, we adopt the *Bilinear* and the precision is set to $\frac{1}{4}$, thereby requiring 135×H×W FLOPs for a 3×H×W RGB input. For the block-based ME for LMV, the search range of hexagonal search is set to 64 and the search block size varies from 64×64 to 4×4, thereby requiring 2587.18×H×W FLOPs for a 3×H×W RGB input. For the HMV generation, according to equation (6) in Section III-C.2, $8 \times H \times W$ FLOPs are performed for HMV generated from LMV in total. Therefore, in once ME, **2730.18×H×W FLOPs** are required for a 3×H×W RGB input in total (upper limit). Note that the variable search block sizes are not tried for each block, early termination algorithm is designed in the hand-crafted ME.

2) TWT & ITWT: It consists of two stages: wavelet transform and selection of different TWT types. First, for the wavelet transform, three types of TWT (temporal forward/backward Haar wavelet transform, temporal CDF 5/3 wavelet transform) are performed on the $3 \times H \times W$ RGB input frames, respectively. And the inverse transform takes the same amount of computational complexity as the forward transform, thereby requiring 60×H×W FLOPs for the 3×H×W RGB input frames. Second, for the selection of different TWT types for each pixel, the high-frequency energy computation and the different type's comparison require 17×H×W FLOPs. Therefore, in once TWT & ITWT, **77×H×W FLOPs** are required for the 3×H×W RGB input frames in total.

3) Enhancement of Low-Frequency Frames: The enhancement of low-frequency frames utilizes the enhancement network illustrated in Fig. 5, thereby requiring **335,232×H×W FLOPs** for a 3×H×W input in total.

TABLE IV
THE ΔLPIPS, ΔMUSIQ AND ΔCLIP-IQA BETWEEN COMPRESSED VIDEOS AND VIDEOS ENHANCED BY DIFFERENT LEVEL TWT-BASED METHODS ON THE TEST SEQUENCES OF JCT-VC AND THE FLOPS DENOTES THE COMPUTATIONAL COMPLEXITY OF ENHANCING THIRTY-TWO FRAMES WITH A RESOLUTION OF 3 × 512 × 512

TWT Decomposed Level	0	1	2	3	4	5
ΔLPIPS	-0.069	-0.068	-0.070	-0.067	-0.064	-0.056
ΔMUSIQ [56]	5.54	4.63	4.24	4.00	3.47	3.19
ΔCLIP-IQA [57]	0.155	0.152	0.183	0.177	0.161	0.157
FLOPs	2812G	1429G	737G	392G	219G	132G

In general, 4-level TWT-based perceptual quality enhancement method is adopted in our framework, and sixteen frames are enhanced simultaneously. Thirty times MEs, fifteen times TWT & TWIT, and once low-frequency frame enhancement are required in total. For the input of sixteen 3×H×W frames, **418292.4×H×W FLOPs** are required in total. For image/single-frame quality enhancement methods which enhance each frame individually, the enhancement network is invoked sixteen times for the input of sixteen 3×H×W frames. If the enhancement network is the same as that in our method, 5,363,712×H×W FLOPs are required. Our 4-level TWT-based method can save nearly 13× the amount of computation. Furthermore, compared with multi-frame assisted quality enhancement methods that employ additional deep learning-based motion alignment, our method can save more amount of computation.

VI. ADDITIONAL ANALYSIS

A. The Impact of Different Level TWT

In this subsection, we analyze the impact of different levels of TWT. For different level TWT-based enhancement methods, Table IV presents the results of ΔLPIPS, ΔMUSIQ [56], and ΔCLIP-IQA [57] between compressed videos and enhanced videos on the JCT-VC test set at QP = 37, along with the FLOPs required for enhancing thirty-two frames with a resolution of 3 × 512 × 512. Note that the decreased LPIPS, increased MUSIQ, and CLIP-IQA indicate better perceptual quality. In Table IV, the 0-level TWT represents that no TWT is employed in our method, which is the same as image/single-frame quality enhancement methods. The 5-level TWT-based enhancement method transforms thirty-two compressed frames into one temporal low-frequency frame and thirty-one high-frequency frames, and only the low-frequency frame is enhanced through the enhancement network. Note that we train the enhancement network of each decomposed level on the temporal low-frequency frames generated by the corresponding decomposed TWT level.

For the 0-level TWT-based method, it achieves a superior subjective quality but with heavy computational complexity. For the 5-level TWT-based method, it can greatly reduce computational complexity, but the performance of quality enhancement is slightly poor. From the results of the three

TABLE V
THE AVERAGE ΔLPIPS, ΔMUSIQ AND ΔCLIP-IQA BETWEEN COMPRESSED VIDEOS AND VIDEOS ENHANCED BY DIFFERENT LEVEL TWT-BASED METHODS WHICH ENHANCES HIGH- AND LOW-FREQUENCY FRAMES SIMULTANEOUSLY

TWT Decomposed Level	0	1	2	3	4	5
ΔLPIPS	-0.069	-0.068	-0.070	-0.066	-0.066	-0.062
ΔMUSIQ [56]	5.54	4.52	3.99	3.30	2.92	2.63
ΔCLIP-IQA [57]	0.155	0.141	0.165	0.176	0.142	0.136

metrics as a whole in Table IV, it demonstrates the trade-off between computational complexity and the performance of quality enhancement that the increase of decomposed levels leads to low performance with low computational complexity. In detail, the number of low-frequency frames decreases with the increase of decomposed levels, and the low-frequency frames can only reflect the less main content of video frames. Therefore, for the enhancement of low-frequency frames, less main content of video frames can be enhanced, and the performance is limited. Therefore, in our default setting, we compromise to consider the trade-off, and adopt the 4-level TWT in our method.

B. The Performance of Enhancing Both High- and Low-Frequency Frames

In this subsection, we present the performance of enhancing temporal high- and low-frequency frames simultaneously. The enhancement network structure of high-frequency frames is the same as that of low-frequency frames, which is shown in Fig. 5. However, the training set is replaced from low-frequency frames with high-frequency frames. On the test set of JCT-VC [11], we conduct the experiment of enhancing both high- and low-frequency frames with 0-level to 5-level TWT-based methods, respectively. Table V shows the average ΔLPIPS, ΔMUSIQ [56], and ΔCLIP-IQA between compressed videos and enhanced videos. Compared with Table IV, the results show that the simultaneous enhancement of high- and low-frequency frames may not obviously improve performance and even decrease performance, which verifies that our motivation that exclusively enhances temporal low-frequency frames is an efficient way to enhance the quality.

C. The Performance of Objective Quality Enhancement

In this subsection, we explore whether our method can be applied for objective quality-oriented enhancement on compressed videos. Specifically, we utilize MSE as the loss function to train the enhancement network. We conduct the test experiment on the test set of JCT-VC [11], and Table VI shows the average PSNR of compressed videos, videos enhanced by the 0-level TWT-based method, and videos enhanced by the 4-level TWT-based method. It can be found that both the 0-level TWT-based method and the 4-level TWT-based method can significantly improve the objective quality of compressed videos. However, compared with the 0.47dB PSNR improvement of the 0-level TWT-based method, the 4-level TWT-based method only has 0.3dB improvement, which is 36% lower than the 0-level TWT-based method. This shows that there are still certain challenges to utilizing the TWT in the objective quality enhancement. The reason is that objective quality is susceptible to the accuracy of pixel values, but the TWT-based method does not process pixels of each frame, resulting in the performance of objective quality enhancement being relatively poor. However, for the perceptual quality enhancement, the pixel value of each frame does not need to be consistent with the original uncompressed pixel value, so the performance of perceptual quality enhancement is better.

TABLE VI
THE AVERAGE PSNR RESULT OF VIDEOS WITHOUT ENHANCEMENT, VIDEOS ENHANCED BY THE 0-LEVEL TWT-BASED METHOD, AND VIDEOS ENHANCED BY THE 4-LEVEL TWT-BASED METHOD ON THE TEST SET OF JCT-VC [11]

Methods	No enhancement	0-level TWT	4-level TWT
PSNR	30.33 dB	30.8 dB	30.63 dB

VII. CONCLUSION

In this paper, we propose the perceptual quality enhancement of compressed videos that improves by exclusively enhancing low-frequency content, which exploits the limited computational resource more efficiently. Specifically, we design a TWT-based low-complexity perceptual quality enhancement method. The TWT with hand-crafted ME is utilized to implement the TFA, which transforms the compressed video into temporal high- and low-frequency frames, then an enhancement network is utilized to enhance the low-frequency frames. Finally, TWIT is performed on the temporal high-frequency frames and the enhanced temporal low-frequency frames to generate the enhanced video. Extensive experimental results show that the performance of our method is comparable with the state-of-the-art methods, and our method has significantly lower complexity.

In the future, our method can be further improved in some aspects. (1) The MV in our method is generated by an additional hand-crafted ME module. However, the compressed video bitstream has contained the motion field information, so is it possible to use the MV directly? This has two potential problems. First, the MV is estimated between the compressed reference frame and the original current frame, which is mismatched with compressed frames in our method. Second, the MV is estimated from different direction reference frames and different distances reference frames, it is difficult to use in our method. (2) From Section V-D, we find that the DNN-based enhancement network occupies most of the computation. To further reduce the computational complexity, we may try to design a more efficient and effective perceptual quality enhancement network with the help of some advanced techniques, such as knowledge distillation.

REFERENCES

[1] Y. Dai, D. Liu, and F. Wu, "A convolutional neural network approach for post-processing in HEVC intra coding," in *Proc. Int. Conf. Multimedia Model. (MMM)*. Cham, Switzerland: Springer, 2017, pp. 28–39.

[2] R. Yang, M. Xu, and Z. Wang, "Decoder-side HEVC quality enhancement with scalable convolutional neural network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 817–822.

[3] R. Yang, M. Xu, Z. Wang, and T. Li, "Multi-frame quality enhancement for compressed video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6664–6673.

[4] Z. Guan, Q. Xing, M. Xu, R. Yang, T. Liu, and Z. Wang, "MFQE 2.0: A new approach for multi-frame quality enhancement on compressed video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 949–963, Mar. 2021.

[5] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6228–6237.

[6] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Deep generative adversarial compression artifact removal," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4836–4845.

[7] L. Galteri, L. Seidenari, M. Bertini, and A. D. Bimbo, "Deep universal generative adversarial compression artifact removal," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2131–2145, Aug. 2019.

[8] D. Chun, T. S. Kim, K. Lee, and H.-J. Lee, "Compressed video restoration using a generative adversarial network for subjective quality enhancement," *IEIE Trans. Smart Process. Comput.*, vol. 9, no. 1, pp. 1–6, Feb. 2020.

[9] J. Wang, X. Deng, M. Xu, C. Chen, and Y. Song, "Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 405–421.

[10] J. Wang, M. Xu, X. Deng, L. Shen, and Y. Song, "MW-GAN+ for perceptual quality enhancement on compressed video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4224–4237, Jul. 2022.

[11] J.-R. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, and T. Wiegand, "Comparison of the coding efficiency of video coding standards—Including high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[12] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 576–584.

[13] Q. Xing, M. Xu, T. Li, and Z. Guan, "Early exit or not: Resource-efficient blind quality enhancement for compressed images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 275–292.

[14] J. He, X. He, M. Zhang, S. Xiong, and H. Chen, "Deep dual-domain semi-blind network for compressed image quality enhancement," *Knowl.-Based Syst.*, vol. 238, Feb. 2022, Art. no. 107870.

[15] Y. Dai, D. Liu, Z.-J. Zha, and F. Wu, "A CNN-based in-loop filter with CU classification for HEVC," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2018, pp. 1–4.

[16] A. Norkin et al., "HEVC deblocking filter," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1746–1754, Dec. 2012.

[17] C.-M. Fu et al., "Sample adaptive offset in the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1755–1764, Dec. 2012.

[18] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[19] T. Wang, M. Chen, and H. Chao, "A novel deep learning-based method of improving coding efficiency from the decoder-end for HEVC," in *Proc. Data Compress. Conf. (DCC)*, Apr. 2017, pp. 410–419.

[20] J. Deng, L. Wang, S. Pu, and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 10696–10703.

[21] S.-J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[22] C. Dong, H. Ma, D. Liu, and J. W. Woods, "Wavelet-based learned scalable video coding," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 3190–3194.

[23] S.-T. Hsiang, "Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," in *Proc. Data Compress. Conf. (DCC)*, 2001, pp. 83–92.

[24] Y. Wu, K. Hanke, T. Rusert, and J. W. Woods, "Enhanced MC-EZBC scalable video coder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 10, pp. 1432–1436, Oct. 2008.

[25] Y. Chen and G. Liu, "Content adaptive Lagrange multiplier selection for rate-distortion optimization in 3-D wavelet-based scalable video coding," *Entropy*, vol. 20, no. 3, p. 181, Mar. 2018.

[26] H. Ma, D. Liu, R. Xiong, and F. Wu, "iWave: CNN-based wavelet-like transform for image compression," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1667–1679, Jul. 2020.

[27] H. Ma, D. Liu, N. Yan, H. Li, and F. Wu, "End-to-end optimized versatile image compression with wavelet-like transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1247–1263, Mar. 2022.

[28] K. Onthriar, K. K. Loo, and Z. Xue, "Performance comparison of emerging dirac video codec with H.264/AV," in *Proc. Int. Conf. Digit. Telecommun.*, 2006, p. 22.

[29] D. T. Nam, G. Y. Gryzov, A. V. Dvorkovich, and V. P. Dvorkovich, "Nonlinear quantization method for wavelet-based video codec," in *Proc. Eng. Telecommun. (EnT-MIPT)*, 2018, pp. 25–29.

[30] K. Bystrov, A. Dvorkovich, V. Dvorkovich, and G. Gryzov, "Usage of video codec based on multichannel wavelet decomposition in video streaming telecommunication systems," in *Distributed Computer and Communication Networks (DCCN)*. Berlin, Germany: Springer, 2017, pp. 108–119.

[31] D. Jin, J. Lei, B. Peng, Z. Pan, L. Li, and N. Ling, "Learned video compression with efficient temporal context learning," *IEEE Trans. Image Process.*, vol. 32, pp. 3188–3198, 2023.

[32] A. Meyer, F. Brand, and A. Kaup, "Learned wavelet video coding using motion compensated temporal filtering," 2023, *arXiv:2305.16211*.

[33] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2720–2729.

[34] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.

[35] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Glasgow, U.K.: Springer, 2020, pp. 402–419.

[36] N. Al-Najdawi, M. Noor Al-Najdawi, and S. Tedmori, "Employing a novel cross-diamond search in a modified hierarchical search motion estimation algorithm for video compression," *Inf. Sci.*, vol. 268, pp. 425–435, Jun. 2014.

[37] B. Furht, J. Greenberg, and R. Westwater, *Motion Estimation Algorithms for Video Compression*, vol. 379. Berlin, Germany: Springer, 2012.

[38] J.-L. Lin, Y.-W. Chen, Y.-W. Huang, and S.-M. Lei, "Motion vector coding in the HEVC standard," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 6, pp. 957–968, Dec. 2013.

[39] W.-J. Chien et al., "Motion vector coding and block merging in the versatile video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3848–3861, Oct. 2021.

[40] B. Bross et al., "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.

[41] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[42] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 214–223.

[43] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.

[44] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 63–79.

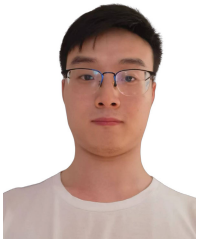
[45] A. Lucas, S. López-Tapia, R. Molina, and A. K. Katsaggelos, "Generative adversarial networks and perceptual losses for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3312–3327, Jul. 2019.

[46] H. Ma, D. Liu, and F. Wu, "Rectified Wasserstein generative adversarial networks for perceptual image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3648–3663, Mar. 2023.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[48] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2019.

[49] Y. Blau, R. Mechrez, R. Timofte, T. Michaeli, and L. Zelnik-Manor, "The 2018 PIRM challenge on perceptual image super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 334–355.
[50] X. Xu, S. Liu, and Z. Li, "Tencent video dataset (TVD): A video dataset for learning-based visual data compression and analysis," 2021, *arXiv:2105.05961*.
[51] D. Ma, F. Zhang, and D. R. Bull, "BVI-DVC: A training database for deep video compression," *IEEE Trans. Multimedia*, vol. 24, pp. 3847–3858, 2022.
[52] F. Bossen, *Common HM Test Conditions and Software Reference Configurations*, document JCTVC-L1100, presented at the 12th JCT-VC Meeting, Geneva, Switzerland, Jan. 2013.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[54] T. Tieleman and G. Hinton, "RMSprop: Divide the gradient by a running average of its recent magnitude," *COURSERA, Neural Netw. Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.
[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[56] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5128–5137.
[57] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 37, 2023, pp. 2555–2563.



Cunhui Dong received the B.S. degree in automation and the M.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2020 and 2023, respectively. His research interests include image/video coding and machine learning.

Haichuan Ma received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 2017, and the Ph.D. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2022. His research interests include image/video coding, signal processing, and machine learning.

Zhuoyuan Li (Student Member, IEEE) received the B.S. degree in communication engineering from Southwest Jiaotong University, Chengdu, China, in 2020. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China. His research interests include video coding and processing.

Li Li (Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2011 and 2016, respectively.

He was a Visiting Assistant Professor with the University of Missouri–Kansas City from 2016 to 2020. He joined the Department of Electronic Engineering and Information Science, USTC, as a Research Fellow, in 2020, and became a Professor in 2022. His research interests include image/video/point cloud coding and processing. He received the Multimedia Rising Star Award from the 2023 IEEE International Conference on Multimedia and Expo (ICME). He received the Best 10% Paper Award from the 2016 IEEE Visual Communications and Image Processing (VCIP) and the 2019 IEEE International Conference on Image Processing (ICIP).

Dong Liu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively.

He was a member of Research Staff with the Nokia Research Center, Beijing, China, from 2009 to 2012. He joined USTC, as an Associate Professor, in 2012, and became a Professor in 2020. He has authored or coauthored more than 200 papers in international journals and conferences. He has more than 30 granted patents. His research interests include image and video processing, coding, analysis, and data mining. He had served as an Organizing Committee Member for VCIP 2022, ChinaMM 2022, and ICME 2021. He is a Senior Member of CCF and CSIG and an elected member of MSA-TC of the IEEE CAS Society. He received the 2009 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award, the VCIP 2016 Best 10% Paper Award, and the ISCAS 2022 Grand Challenge Top Creativity Paper Award. He serves as the Chair for the IEEE 1857.11 Standard Working Subgroup (also known as the Future Video Coding Study Group). He serves as an Associate Editor for *Frontiers in Signal Processing*.