Enlarged Motion-Aware and Frequency-Aware Network for Compressed Video Artifact Reduction

Wang Liu, Wei Gao, *Senior Member, IEEE*, Ge Li, *Member, IEEE*, Siwei Ma, *Fellow, IEEE*, Tiesong Zhao, *Senior Member, IEEE*, and Hui Yuan, *Senior Member, IEEE*

***Abstract*— Making full use of spatial-temporal information is the key factor for removing compressed video artifacts. Recently, many deep learning-based compression artifact reduction methods have emerged. Among them, a series of methods based on deformable convolution have shown excellent capabilities in spatio-temporal feature extraction. However, local deformable offset prediction and pixel-wise inter-frame feature alignment in the unidirectional form limit the full utilization of temporal features in the existing method. Additionally, compressed video shows inconsistent degrees of distortion on different frequency components, and their restoration difficulty is also nonuniform. For the above problems presented by existing methods, we propose an *enlarged motion-aware and frequency-aware network* (EMAFA) to further extract spatio-temporal information and enhance information of different frequency components. To perceive different degrees of motion artifacts between compressed frames as accurately as possible, we design a bidirectional dense propagation pattern with *pixel-wise and patch-wise deformable convolution* (PIPA) module in the feature domain. In addition, we propose a *multi-scale atrous deformable alignment* (MSADA) module to enrich spatio-temporal features in image domain. Moreover, we design a *multi-direction frequency enhancement* (MDFE) module with multiple direction convolution to enhance the features of different frequency components. The experimental results show that the proposed method performs better than the state-of-the-art methods in both objective evaluation and visual perception experience. Supplementary experiments for Internet Streamed Video with hybrid-distortion demonstrate that our method also exhibits considerable generalizability for quality enhancement.**

***Index Terms*— Video compression, compressed video artifact reduction, video quality enhancement.**

I. Introduction

THE demand for information is increasing with the continuous development of computer network technology and multimedia. In particular, multimedia technology has become the most important carrier for people to obtain information. As one of the most important links in multimedia technology, video compression technology has also developed rapidly. Developing video compression techniques to significantly reduce the bit rate is essential to transmit video under limited bandwidth. However, lossy compression algorithms such as H.265/HEVC [1] inevitably introduce various artifacts, resulting in severe video quality degradation. In general, post-processing on the decoding end is usually employed to improve the compressed video quality after the video stream is decoded.

Compression artifact removal methods have been extensively studied in recent years. The basic goal of the compression artifact removal task is to improve compressed sequences quality at the decoder. Some studies [2], [3], [4] have shown that joint optimization methods at the encoder can improve compressed video quality, which is beyond the scope of our work. Other research has focuses on enhancing the compression quality of the perceptual experience [5]. Compared with the traditional artifact removal technologies, the deep learning-based methods with better performance have received increasing attention from both academia and industry. Deep learning-based methods transform the compressed artifact removal tasks into an end-to-end prediction processes [6], [7], [8], [9], [10], [11], [12], [13], leveraging the powerful feature extraction and feature fusion capabilities. The learning-based methods for video compression artifact removal roughly consist of two steps: feature alignment and feature fusion. The former is used to estimate the spatio-temporal relationship of the compressed video, while the latter fuses the aligned features. Learning-based methods can be divided into four categories based on differences in the feature alignment process: explicit or implicit optical flow estimation, deformable convolution and self-attention mechanisms. The feature fusion process can be classified as unidirectional or multi-directional. Most methods tend to

Manuscript received 8 August 2023; revised 17 April 2024; accepted 20 May 2024. Date of publication 28 May 2024; date of current version 30 October 2024. This work was supported in part by The Major Key Project of PCL under Grant PCL2024A02, in part by the Natural Science Foundation of China under Grant 62271013 and Grant 62031013, in part by Guangdong Province Pearl River Talent Program under Grant 2021QN020708, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515010155, in part by Shenzhen Science and Technology Program under Grant JCYJ20230807120808017, and in part bythe CAAI-MindSpore Open Fund, developed on OpenI Community under Grant CAAIXSJLJJ-2023-MindSpore07. This article was recommended by Associate Editor Q. Xu. *(Corresponding author: Wei Gao.)*

Wang Liu is with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: liuwang@stu.pku.edu.cn).

Wei Gao and Ge Li are with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen 518055, China (e-mail: gaowei262@pku.edu.cn; geli@ece.pku.edu.cn).

Siwei Ma is with the National Engineering Research Center of Visual Technology, School of Computer Science, Peking University, Beijing 100871, China (e-mail: swma@pku.edu.cn).

Tiesong Zhao is with Fujian Key Laboratory for Intelligent Processing and Wireless Transmission of Media Information, Fuzhou University, Fuzhou, Fujian 350116, China (e-mail: t.zhao@fzu.edu.cn).

Hui Yuan is with the School of Control Science and Engineering, Shandong University, Jinan 250061, China (e-mail: huiyuan@sdu.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TCSVT.2024.3406425.

Digital Object Identifier 10.1109/TCSVT.2024.3406425

migrate the spatio-temporal feature extraction strategies used in raw video processing tasks to compressed video, such as video super-resolution or video interpolation. Therefore, they seldom analyze the impact of the compression process on spatio-temporal feature extraction, and consider improving the ability of the model to remove artifacts according to the characteristics of video compression.

Deformable convolution-based methods can extract richer spatio-temporal features through dynamic deformable offsets and outperforms other methods in terms of both model performance and complexity. However, most methods based on fixed convolution receptive fields fail to extract spatio-temporal correspondence for long-range fast motion. As shown in Fig. 1, deformable convolution with multiple dilation factors in our method can provide richer and more diverse offsets. Compared with regular deformable convolution, our proposed multi-scale atrous deformable alignment (MSADA) module can further remove the compression artifacts.

In addition, residual quantization and quantization parameter adjustment in video compression algorithms inevitably result in inconsistent quality loss for both intra frame and inter frame. From the perspective of the frequency domain, most existing restoration methods tend to focus on high-frequency features. For example, we take the sequence *BasketballPass* from the standard testing sequence [1]. As pioneering work based on the deformable convolution framework, the inter-frame alignment in the STDF model [10] has been widely adopted by most existing methods. Fig. 2 shows that the sequence is compressed and then enhanced by the STDF. Compared with the low-frequency region, the objective metric of the high-frequency region has been significantly improved. In contrast, our method performs well on non-high-frequency feature enhancement.

In summary, according to the compression distortion characteristics and the defects in the above methods, we propose an enlarged motion-aware and frequency-aware network (EMAFA) to extract spatio-temporal information effectively and enhance information with different frequency components. The main contributions of our work are as follows:

- The fixed deformable receptive field adopted by most existing methods fails to extract spatio-temporal information sufficiently. To enlarge the motion perception capability of deformable convolution, we design a multi-scale atrous deformable alignment (MSADA) module with flexible sampling positions to accurately perceive spatio-temporal information from long-range displacement motion.
- Existing methods ignore the fact that video compression algorithms based on block-partitioning inevitably result in inconsistent motion distributions between frames. We propose a pixel-wise and patch-wise deformable convolution (PIPA) module to extract non-local spatio-temporal features to perceive different degrees of motion artifacts.
- Residual quantization and quantization parameter adjustment in video compression irreversibly lead to nonuniform quality distribution of compressed video. Unlike existing methods that roughly classify features into high and low frequencies, we develop a plug and play multi-direction frequency enhancement (MDFE) module with multiple-direction convolution (MDC) to enhance the perception of features with different frequencies simultaneously.

The remainder of this paper is organized as follows: Section II reviews the previous related works. Section III presents our proposed framework. The experimental results and corresponding analysis are described in Section IV. Section V concludes our work.

II. Related Work

The compressed video enhancement method based on deep learning generally processes the compressed frame in a post-processing manner. Compared with traditional compression restoration methods [14], [15], CNN-based models make the compressed video restoration task an end-to-end and learnable predictive enhancement process. CNN-based methods can be divided into two main categories: single-frame enhancement methods [6], [16], [17], [18], [19] and multi-frame enhancement methods, respectively. In particular, multi-frame enhancement methods can be divided into methods based on explicit or implicit optical flow estimation (OFE) [7], [8], [9], [20], [21], [22], methods based on deformable convolution (DCN) [10], [11], [23], [24], [25] and methods based on self attention [26]. OFE-based methods mainly predict temporal information through optical flow calculations between adjacent frames, and finally realize the prediction and enhancement of the target frame. DCN-based methods extract richer spatio-temporal features through dynamic deformable offsets and outperform OFE-based methods in terms of performance and robustness. Due to their excellent nonlocal feature extraction ability, video enhancement methods based on self-attention mechanisms have recently emerged [26]. However, its slow inference speed and weak locality limit its application to fine-grained low-level enhancement tasks.

In this section, we briefly describe the related work on deformable convolution and existing compressed video artifact reduction methods. Moreover, we systematically analyze the spatio-temporal feature extraction problem of compressed video in terms of commonality and particularity, which differs from general video processing tasks.

A. Deformable Convolution

Deformable Convolution [28], [29]was first proposed for the object detection task, which makes the sampling location in regular convolution more flexible by introducing a learnable offset. Several studies have shown that deformable convolution performs well in tasks related to video processing, such as video super-resolution [11], [30], video demoiréing [31] and video deblurring [32], [33]. Extracting sufficient temporal information between video frames is always key in the above tasks.

B. Compressed Video Artifact Reduction

To achieve frame alignment and feature fusion in an end-to-end manner, Xue et al. [7] proposed a task-oriented flow

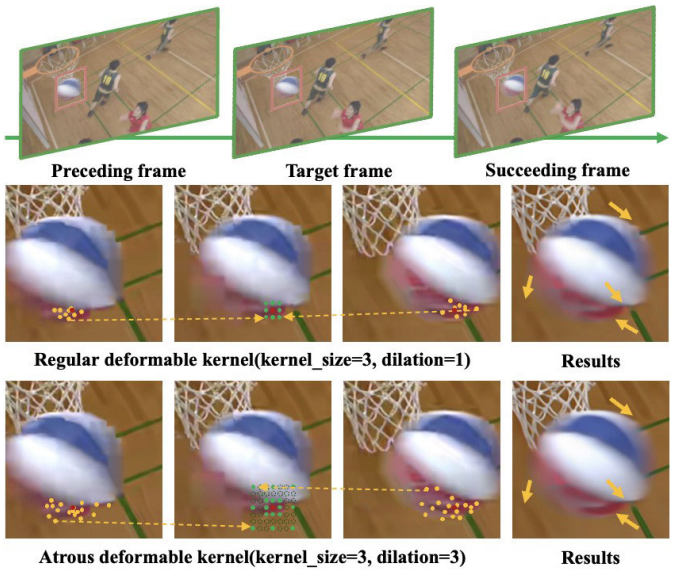


Fig. 1. The illustration of the multi-scale atrous deformable alignment (MSADA) module. Compared with regular deformable convolution, the proposed MSADA module with multiple dilation factors provides rich and diverse offsets for long-range motion displacement. The abundant alignment features integrated by the MSADA module in the image domain can ensure the accuracy of subsequent feature fusion.

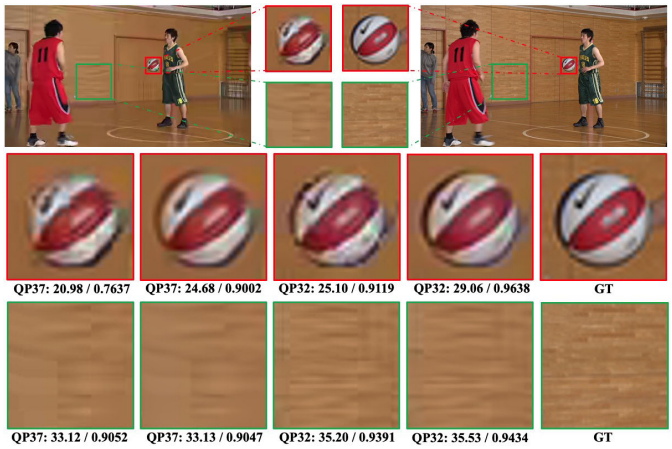


Fig. 2. Artifact level and enhanced result in regions with different frequencies. The second and third rows represent the high-frequency region and the low-frequency region, respectively. The first four columns sequentially represent compressed and enhanced frames under Quantization Parameter $QP = 37$, and compressed and enhanced frames under $QP = 32$. The raw frame is shown in the last column. PSNR (dB) and SSIM [27] are chosen as objective quality evaluation metric.

model (TOFlow) to remove blocking effects. Inter-prediction in video compression leads to severe quality fluctuations, which are ignored by previous methods. Guan et al. [8], [34] designed a multi-frame quality enhancement strategy (MFQE) to improve enhancement performance, which use the high-quality frame (HQF) to improve the low-quality frame (LQF). However, HQF-based methods ignore the fact that some high-quality detail regions still exist in LQF. Hence, Xu et al. [20] proposed the NL-ConvLSTM model, which can capture hidden spatio-temporal information in the adjacent frames. Ding et al. [9] analyzed the quality variation in HQF and proposed the PMVE model to further improve the enhancement quality by adopting a hierarchical prediction enhancement mechanism. Ding et al. [21] designed a block-wise spatio-temporal quality enhancement framework that adaptively utilizes and enhances compressed sub-blocks through temporal and spatial information. Existing methods usually use bilinear interpolation for motion prediction, which introduces additional artifacts and blurring effects. Therefore, Bao et al. designed a spatially variable interpolation kernel during frame alignment and propose the MEMC-Net model [22].

However, compression artifacts inevitably affect the effectiveness of the optical flow estimation process and spatio-temporal features. In particular, it is challenging for OFE-based methods to perform offset compensation explicitly or implicitly for large-displacement motion. Compared with these methods, DCN-based methods have advantages in terms of temporal information diversity and computational overhead. Deng et al. proposed a spatio-temporal defomable convolution network to fuse information for compressed video enhancement [10]. The temporal correlation between adjacent frames obtained through deformable convolution is robust to compression artifacts. However, it is difficult for deformable convolution to address the blur caused by fast-moving objects and compressed artifacts simultaneously. Therefore, Wang et al. [11] designed a pyramid-level deformable alignment unit to handle fast movement separately. Zhao et al. designed a recursive fusion form and deformable attention module for quality enhancement [23]. For the problem of inconsistent artifact levels in the frequency domain, Peng et al. [24] built an omni-frequency adaptive enhancement (OFAE) block to improve the quality of the detailed region. Luo et al. [25] established a multi-scale alignment unit to produce more accurate deformble offsets. Wang et al. [35] proposed a spatio-temporal information balance network to reduce alignment errors and obtain a balance between temporal and spatial information. Since most existing methods seldom consider the problem of displacement diversity, Wang et al. [36] proposed a spatio-temporal decomposition network to adaptively fuse the spatial and temporal information.

C. The Particularity of Spatio-Temporal Correspondence for Compressed Video

At present, the main spatio-temporal correspondence extraction methods in deep learning-based video processing include explicit optical flow, long and short-term memory (LSTM), 3D convolution [37], [38], [39], deformable convolution, and self-attention. Despite the fact that the training process is slow, the deformable convolution-based methods stand out in terms of performance, inference speed, and stability. However, regular deformable convolution is not always suitable for spatio-temporal feature extraction in compressed videos. For video super-resolution or video completion tasks, the motion consistency between frames is greater than that of compressed video. Additionally, features at different frequencies have similar distortion distributions for video super-resolution. Hence, existing DCN-based spatio-temporal feature extraction

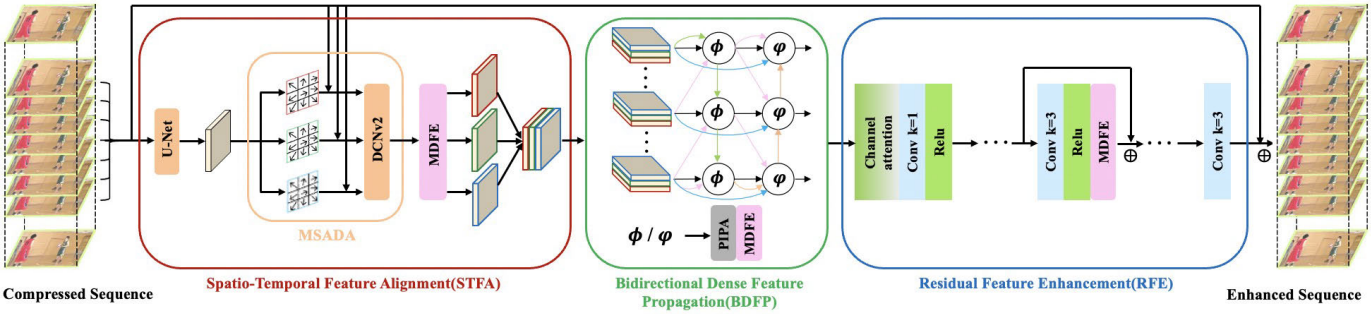


Fig. 3. Overall framework of the proposed EMAFA model. It mainly involves three units: *Spatio-Temporal Feature Alignment* (STFA) unit for coarse feature alignment, *Bidirectional Dense Feature Propagation* (BDFP) unit for fine feature alignment and *Residual Feature Enhancement* (RFE) unit for residual learning. The STFA unit involves multi-scale atrous offset prediction, deformable convolution process and multi-direction frequency enhancement (MDFE) module. The first two constitute the MSADA module. The BDFP unit includes two dense feature propagation layers ϕ and φ. Each layer contains a PIPA module and an MDFE module. The RFE unit consists of a channel attention layer and several residual blocks with the MDFE module.

strategies need further optimization to adapt to compressed video. The main explanations are as follows:

- Block partitions of different sizes and block-level motion estimation errors in video compression irreversibly cause inconsistency and non-locality in the distribution of spatio-temporal correspondence. It is difficult to capture long-range fast motion with a fixed convolutional receptive field at a single resolution. The MSADA module proposed in our method is dedicated to more accurately perceiving spatio-temporal motion for compressed video.
- Furthermore, most existing methods are limited to the local spatio-temporal correlation of compressed video. The PIPA module proposed in our method can receive the global motion between frames.

III. Method

A. Problem Formulation and Framework Overview

Given a compressed video sequence with N frames $\left\{I_i^{LQ} \mid i = 1, 2, \cdots, N\right\}$, the goal of compressed artifact reduction is to remove the distortion caused by motion estimation and residual quantization, and ultimately improve video quality. Different from compressed image restoration, compressed video artifact reduction usually uses a multi-frame enhancement framework to exploit temporal information. According to the output form, the multi-frame enhancement framework can be classified into two types: single-frame output and multi-frame output. For a clip with $2R + 1$ consecutive frames $\left\{I_i^{LQ} \mid i = -R, -R+1, \cdots, R\right\}$ in the above sequence, the goal of artifact reduction is to obtain an intermediate frame output I_0^{HQ} or a sequence output $\left\{I_i^{HQ} \mid i = -R, -R+1, \cdots, R\right\}$ with high quality.

The overall framework of our proposed method is presented in Fig. 3, which mainly consists of three units: *spatio-temporal feature alignment* (STFA) unit, *bidirectional dense feature propagation* (BDFP) unit and *residual feature enhancement* (RFE) unit. The STFA unit is dedicated to extracting spatio-temporal features in sequence. It consists of two proposed modules: *multi-Scale atrous deformable alignment* (MSADA) and *multi-direction frequency enhancement* (MDFE).

First, the MSADA module extracts coarse alignment features for each input frame in the image domain. With the help of the flexible receptive field in the MSADA module, deformable convolution can capture richer spatio-temporal information with fewer parameters, especially for large displacement motion. Since high-frequency features and non-high-frequency features have different degrees of artifact in compressed video, the MDFE module is designed to enhance features with different frequency through a divide-and-conquer strategy.

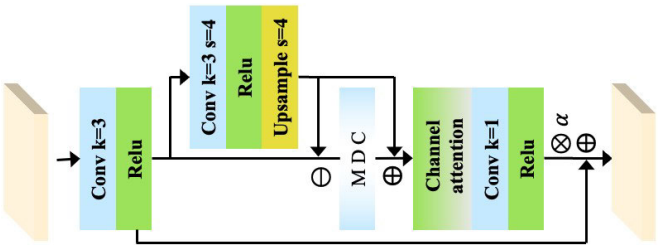


Fig. 4. The architecture of the multi-direction frequency enhancement (MDFE) module.

After that, the BDFP unit allows the coarse features obtained by MSADA to propagate and merge in a multi-layer form. Bidirectional information propagation within layers and dense connections between layers in the feature domain will further refine the existing coarse features. Moreover, a novel *Pixel-Wise and Patch-Wise Deformable Alignment* (PIPA) block is proposed to obtain local and non-local spatial features for alleviating motion artifacts based on coding unit partitioning.

Finally, the coarse features obtained by the STFA unit and the refined features in all layers from the BDFP unit are concatenated to the RFE unit. The RFE unit focuses on residual estimation between the compressed frame and raw frame, which includes a channel attention layer and several residual blocks with the MDFE module.

B. Spatio-Temporal Feature Alignment Unit

For a clip of seven consecutive frames $\left\{I_{-3}^{LQ}, I_{-2}^{LQ}, \cdots, I_3^{LQ}\right\}$, i.e.$R = 3$. To quickly extract deep features of all frames at different scales, we directly adopt the U-Net architecture in STDF [10] and obtain the corresponding depth features $\left\{F_{-3}^{LQ}, F_{-2}^{LQ}, \cdots, F_3^{LQ}\right\}$ for all

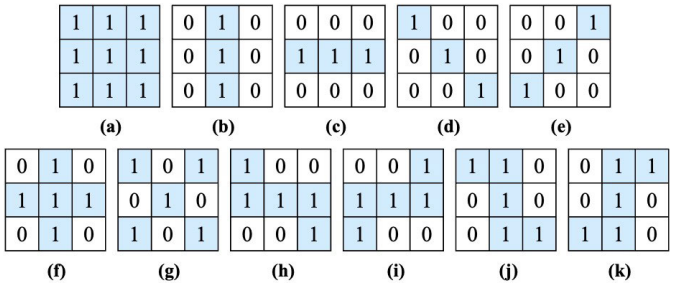


Fig. 5. The illustration of the multiple direction convolution (MDC) module. It separates and extracts different frequency features by controlling the specific sampling position of the convolution receptive field, including high-frequency features and non-high-frequency features. The sampling position is generated by applying the corresponding matrix mask to the convolution kernel.

frames. For the convenience of description, we set I_0^{LQ} as the target frame and consider how to extract spatio-temporal information with feature F_0^{LQ}.

C. Multi-Scale Atrous Deformable Alignment

Limited by the fixed-size deformable convolution kernel, it is difficult for existing methods to extract temporal information effectively. Most existing methods are ineffective when encountering long-range motion displacement or non-adjacent frames. Expanding the convolution kernel size to extract non-local features in STDR [25] will cause a sharp increase in the model parameter size. For the above deficiency, we propose the MSADA module to extract spatio-temporal information by setting multiple dilation factors. The regular 3×3 convolution process is described as follows:

$$F_0^a(p) = \sum_{t=-3}^{3} \sum_{k=1}^{9} W_{t,k} \cdot I_t^{LQ}(p + p_k), \tag{1}$$

where $F_0^a(p)$ is the aligned feature on point p of target frame I_0^{LQ}, t and k denote the indices of the reference frame and sampling position, respectively. Limited to a fixed sampling position, it is difficult for regular convolution to deal with video tasks with temporal information. To effectively extract temporal features, deformable convolution is designed to select the receptive range dynamically in [10]. The dynamic sampling process is achieved through a learnable offset, which is expressed as follows:

$$p_k \leftarrow p_k + (\delta^h_{(t,p),k}, \delta^v_{(t,p),k}) \tag{2}$$

where $\delta^h_{(t,p),k}, \delta^v_{(t,p),k} \in R$ indicate the horizontal and vertical position offset of the k-th sampling point for the center point p in the t-th frame, respectively.

For the regular 3×3 convolution with dilation factor $d = 1$, sampling position $p_k \in \{(-1,-1), (-1,0), \cdots, (1,1)\}$. The unit size used in the block partition-based video codec framework ranges from 8×8 to 64×64, which accounts for blocking artifacts. Our method adjusts the receptive range of convolution to cover the smallest block boundary to alleviate the blocking effect in compression artifacts. We choose dilation factors $d = 3$ and $d = 5$ to perform non-local and long-range feature extraction with a limited parameter size. In this case, sampling position changes to $p_k \in \{(-3,-3), (-3,0), \cdots, (3,3)\}$ and $p_k \in \{(-5,-5), (-5,0), \cdots, (5,5)\}$, respectively. The final key point is how to learn the offset suitable for sampling positions of different scales. As shown in Fig. 3, we apply different 3×3 convolution layers to the feature F_0^{LQ} to adaptively learn the offset for the above three types of sampling locations. With the help of the deformable convolution module in MMDetection toolbox [40], the spatio-temporal aligned feature F_0^a is obtained. Similarly, a set of aligned features $\{F_i^a \mid i = -3, -2, \cdots, 3\}$ are derived by replacing the target frame and reference frame through parameter sharing.

D. Multi-Direction Frequency Enhancement

In Section I, we have analyzed the fact that compression artifacts at different frequencies exhibit different degrees. Existing methods often focus on enhancing high-frequency features or using loss functions related to the frequency domain, directly resulting in the absence of mid-frequency and low-frequency features. Extracting and enhancing features of different frequencies simultaneously is crucial for removing compression artifacts. Inspired by the fact that enhanced asymmetric convolution blocks [41], [42] can be embedded in image processing tasks, we design a multi-direction frequency enhancement (MDFE) module with the multiple direction convolution (MDC) manner to extract and enhance features with different frequencies separately. The experimental results show that multi-directional convolution has advantages for inconsistent compression artifacts with different frequencies.

As Fig. 4 shows, a 3×3 convolutional layer and a ReLU layer are first applied to the input feature. Then, a 3×3 convolutional layer with stride 4 and bilinear upsampling layer are used to filter out noise. After that, the non-noise features are obtained by subtracting noise. For non-noise features, we utilize the MDC module to extract features with more detailed frequency levels. As shown in Fig. 5, several 3×3 convolutional kernels with specific sampling positions are designed, and we obtain several fine-grained features with different frequencies. To enhance the interaction of features between channels and cut down parameter redundancy, the enhanced features output by the MDC module are further concatenated by a channel attention layer, a 1×1 convolutional layer, and a ReLU layer. Next, a learnable weight parameter is applied to the enhanced features to fit the residual features adaptively followed by pointwise addition to the input feature. Finally, the enhanced features $\{F_i^c \mid i = -3, -2, \cdots, 3\}$ refined by the MDFE module for all frames will participate in the next feature alignment. The MDFE module can be configured as a plug-and-play block for common image enhancement tasks.

Remarks: A convolutional layer with learning parameters and a parameterless upsampling layer are applied to the features before the MDC module to enable the network to adaptively extract noise. Noise arises from the blockiness crossover phenomenon during video compression, and they should not be included in the enhancement process. Then, the subtraction operation is used to filter out noise, and non-noise features are obtained and subsequently enhanced by the

MDC module. The motivation for adding noise again is that it contains some low-frequency features. The above process can gradually dilute or reduce the proportion of noise in the features, especially in the first few layers of the network. In fact, separate extraction layers for noise can be removed in the last few layers of the network.

E. Bidirectional Dense Feature Propagation Unit

The aligned features output by the STFA unit are too coarse to guarantee the accuracy of spatio-temporal correspondence. The reasons why the features from the STFA unit cannot directly participate in subsequent residual feature enhancement are as follows: first, feature extraction and alignment in STFA are performed in the image domain, which suffers from the low quality of the reference frame itself. Second, the offset estimation with compressed artifacts in STFA is a suboptimal result for the deformable convolution. Third, the index distance between the target and the adjacent frame makes the aligned features inconsistent. Therefore, a bidirectional and progressive structure was proposed to overcome the above difficulties in [24] and [43]. However, only pixel-level feature alignment in the feature domain occurs in existing methods. Moreover, the propagation process of a long sequence will inevitably exacerbate error accumulation. In view of the shortcomings of existing methods, we propose a bidirectional dense feature propagation (BDFP) unit equipped with a pixel-wise and patch-wise deformable Alignment (PIPA) block.

Specifically, given a set of coarse aligned features $\{F_i^c \mid i = -3, -2, \cdots, 3\}$ for all frames, the refined features $\{F_i^r \mid i = -3, -2, \cdots, 3\}$ are generated through two dense feature propagation layers ϕ and φ. The current i-th feature F_i^l in layer ϕ is derived as follows:

$$F_i^l = \phi(F_i^c) = \psi(Cat(F_i^{d_1}, F_i^{d_2}) + F_i^c \tag{3}$$

where $Cat(\cdot)$ represents the concatenation operation, ψ contains a channel attention layer, a convolution layer and the MDFE module. $F_i^{d_1}$ and $F_i^{d_2}$ are defined as follows:

$$\begin{aligned} F_i^{d_1} &= d_{pi}(Cat(F_{i-1}^c, F_i^c, F_{i+1}^l), \quad \text{if i} = 2, 1, \cdots, -3, \\ &= d_{pi}(Cat(F_{i-1}^c, F_i^c, F_i^c), \quad \text{otherwise}, \end{aligned} \tag{4}$$

$$\begin{aligned} F_i^{d_2} &= d_{pa}(Cat(F_{i-1}^c, F_i^c, F_{i+1}^l), \quad \text{if i} = 2, 1, \cdots, -3, \\ &= d_{pa}(Cat(F_{i-1}^c, F_i^c, F_i^c), \quad \text{otherwise}, \end{aligned} \tag{5}$$

where d_{pi} and d_{pa} are deformable convolution operators in our proposed PIPA module. The operator d_{pi} directly adopts the U-Net architecture in STDF [10] to further extract spatio-temporal information at the pixel-wise level. The accuracy of spatio-temporal information required for the video compression artifact removal task is much greater than that for the video super-resolution. The temporal correlation of the latter is close to local and uniform, while the former suffers from non-uniformity caused by block artifacts and motion estimation. Inspired by patch embedding in transformers [26], features are divided into patches and participate in deformable convolution. Next, the architecture of the PIPA block is stated briefly.

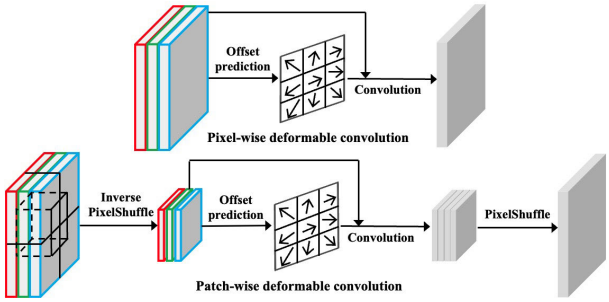


Fig. 6. The illustration of Pixel-Wise and Patch-Wise Deformable Alignment (PIPA) block. It consists of regular pixel-wise deformable convolution and patch-wise deformable convolution.

1) Pixel-Wise and Patch-Wise Deformable Alignment: As shown in Fig. 6, the PIPA block applies deformable convolution to features $F_{i-1}^c, F_i^c, F_{i+1}^{d_2}$ both at the pixel level and patch level. For patch-wise deformable convolution operator d_{pa}, we first use an *Inverse PixelShuffle* layer with a scale factor 4 for patch division, followed by a bottleneck layer with 1×1 convolution layer and a ReLU layer. After that, the deformable offset prediction and feature alignment process is the same as that in d_{pi}. Finally, a *PixelShuffle* layer is used to transform the features to the original resolution. The relevant process is summarized in Equation (4) and Equation (5).

The features $F_i^{d_1}$ and $F_i^{d_2}$ obtained from the PIPA block are further fed into the layer ψ. The refined features F_i^l of the layer ϕ are generated by pointwise addition with F_i^c. The relevant process is summarized in Equation (3). To reduce the error accumulation during feature propagation and increase training process stability, we consider adjusting the direction of propagation and employing skip dense connections in the layer φ. The propagation process is similar to that in layer ϕ, and the specific process is described as follows:

$$F_i^r = \varphi(F_i^l) = \psi(Cat(F_i^{d_1}, F_i^{d_2}, F_i^c) + F_i^l \tag{6}$$

$$\begin{aligned} F_i^{d_1} &= d_{pi}(Cat(F_{i-1}^r, F_i^l, F_{i+1}^l), \quad \text{if i} = -2, -1, \cdots, 3, \\ &= d_{pi}(Cat(F_i^l, F_i^l, F_{i+1}^l), \quad \text{otherwise}, \end{aligned} \tag{7}$$

$$\begin{aligned} F_i^{d_2} &= d_{pa}(Cat(F_{i-1}^r, F_i^l, F_{i+1}^l), \quad \text{if i} = -2, -1, \cdots, 3, \\ &= d_{pa}(Cat(F_i^l, F_i^l, F_{i+1}^l), \quad \text{otherwise}. \end{aligned} \tag{8}$$

F. Residual Feature Enhancement Unit

The coarse alignment features $\{F_i^c\}$ in the STFA unit are promoted by the BDFP unit, which derives the fine alignment features $\{F_i^r\}$. In fact, the spatio-temporal feature extraction and alignment in the above units approximately reconstruct the motion estimation in the video compression process. However, the information loss caused by the quantization residual is still not recovered. Most of the existing methods use several consecutive residual blocks to learn the missing information between the compressed frame and the raw frame. However, minor adjustments of the quantization parameter between frames under the Low Delay P (LDP) configuration leads to non-uniform residual distribution. In this case, we consider embedding the proposed MDFE module into all residual blocks to fit residual information with different frequencies.

As shown in Fig. 3, the propose pesidual feature enhancement (RFE) unit combines a channel attention (CA) layer and several residual blocks with the MDFE module. The widely used CA layer plays an important role in channel information interaction and model parameter size reduction. The output R_i of the RFE unit is assumed to approximate the difference between the compressed frame and the raw frame. The reconstructed frames $\hat{I}_i^{HQ}$ with high quality are calculated by:

$$\hat{I}_i^{HQ} = R_i + I_i^{LQ}, i = -3, -2, \cdots, 3 \tag{9}$$

G. Loss Function

Since all the units in our proposed model are composed of regular convolution and deformable convolution, model parameters are trained end-to-end. According to the settings of most existing models, the Charbonnier et al. [44] loss function is adopted to optimize the model. Assume $\left\{\hat{I}_i^{HQ}\right\}$ and $\left\{\tilde{I}_i^{HQ}\right\}$ are the reconstructed high-quality frame set and the raw high-quality frame set, respectively, and we concatenate them into clips along the temporal axis for training, denoted as $\hat{I}^{HQ}$ and $\tilde{I}^{HQ}$. The loss function is defined as follows:

$$\mathcal{L} = \sqrt{(\tilde{I}^{HQ} - \hat{I}^{HQ})^2 + \varepsilon}, \varepsilon = 1e^{-8}. \tag{10}$$

IV. Experiments

A. Datasets

In our experiments, we choose raw sequences from MFQE2.0 dataset [34]. This dataset includes 108 training sequences collected from Xiph.org and 18 standard testing sequences recommended by the Joint Collaborative Team on Video Coding [1]. The video resolution of the training sequence varies from 352 × 240 to 2560 × 1600. For a fair comparison, all the raw sequences are compressed by the H.265/HEVC reference software HM16.5 under the Low Delay P (LDP) configuration, which is described in detail in MFQE2.0. To evaluate the method performance and compare with other models, all training sequences and testing sequences are compressed at four commonly used Quantization Parameter (QPs), i.e., 22, 27, 32, and 37.

During the training phase, all raw sequences and the corresponding compressed sequences are randomly cropped into 128 × 128 clips from septuplets, which include seven consecutive frames. Unlike other video enhancement tasks, i.e., video interpolation and video super-resolution, compression artifact reduction is subject to the maximum size of the block partition (64 × 64) in the HEVC compression standard. Therefore, 128×128 is usually selected as the sample size to imitate the block prediction process of inter or intra frames as much as possible. Similar to existing methods, data augmentation techniques (i.e., flip or rotation) are introduced to improve the generalization ability. Furthermore, we increase the sampling stride to reduce the similarity of training samples. Compared with the commonly used clips of seven consecutive frames, sampling at intervals of several frames can enrich the inter-frame motion information.

For testing, all test sequences are enhanced frame by frame while maintaining the original resolution. For high-resolution sequences, cropped patches are processed and stitched. As in previous work, we focus on improving the quality of the Y-channel in the YCbCr space. In addition, we choose the increased Peak Signal-to-Noise Ratio (△PSNR) and Structure Similarity (△SSIM) [27] as objective quality metrics to evaluate the improvement in quality.

B. Training Settings

Our method is implemented on the PyTorch platform. With the help of the MMDetection toolbox for deformable convolution [40], the training process is further accelerated. Adam optimizer [45] is adopted with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\varepsilon = 10^{-8}$. The initial learning rate is set to 2×10^{-4} and decays to 1×10^{-4} by a cosine decrease after 80K iterations. The total number of iterations is set to 400K. The batch size is set to 8. In general, the residual between the compressed and raw frames decreases as the quantization step size decreases under the video lossy compression mode. Therefore, the model is first trained with $QP = 37$ and fine-tuned with $QP = 32, 27$, and 22.

C. Comparison With Other Methods

1) Overall Performance: To demonstrate the superiority of the proposed framework in removing compression artifacts, we compare the model with the state-of-the-art methods: AR-CNN [6], DnCNN [19], MFQE2.0 [34], STDF-3L [10], PWSTQ [21], RF-DSTA [23], STDR [25], STDN [36] and STIB [35]. Table. I and Table. II present the quantitative results of accuracy and model complexity, respectively. Most results are taken directly from the original published paper. As shown in Table. I, it shows that our method outperforms all existing methods in terms of both PSNR and SSIM increase under all QPs. In particular, compared with that of STDR, the performance gain reaches 0.13 dB / 0.16 under $QP = 32$ in our method. In Table. II, we also consider the number of reference frames, computational complexity and model size. For a fair comparison, we adjust some methods to achieve similar model sizes by controlling the number of feature channels and reference frames, which are denoted as STDF* [10], BasicVSR++* [43], and OVQE* [24]. Almost all existing multi-frame enhancement methods use seven consecutive frames as input, while OVQE uses 15 consecutive frames in both the training and testing processes. As we know, using more reference frame information can greatly improve the enhancement results, which has been proved by many video processing tasks. We select seven consecutive frames and retrain them using the official code provided to achieve a fair comparison and reduce training time. The retrained models are denoted as OVQE* in Table. II. Compared with STDF*, our method achieves greater quality improvement with fewer model parameters. This demonstrates that the enhanced motion perception and frequency perception proposed in our method is superior to simply enlarging the model size. Compared with BasicVSR++* and OVQE*, our method still achieves the best performance with less computational overhead. Table. III expands the results in Table. II according to the resolution of the video sequence.

TABLE I

OVERALL PERFORMANCE COMPARISON FOR △PSNR (dB) AND △SSIM ($\times 10^{-2}$) OVER STANDARD TEST SEQUENCES AT FOUR QPS. CALCULATE AT QP = 22, 27, 32 AND 37. THE BEST RESULTS IN EACH ROW ARE MARKED IN BOLD

QP	Test sequences		AR-CNN	DnCNN	MFQE2.0	STDF-R3L	PWSTQ	RF-DSTA	STDR	STDN	STIB	Ours
37	Class A	Traffic	0.24 / 0.47	0.24 / 0.57	0.59 / 1.02	0.73 / 1.15	0.64 / 1.04	0.80 / 1.28	0.85 / 1.34	0.83 / 1.32	0.81 / **1.42**	**0.90** / 1.40
	2560x1600	PeopleOnStreet	0.35 / 0.75	0.41 / 0.82	0.92 / 1.57	1.25 / 1.96	1.08 / 1.68	1.44 / 2.22	1.53 / 2.34	1.48 / 2.25	1.41 / 2.32	**1.63 / 2.45**
	Class C	Kimono	0.22 / 0.65	0.24 / 0.75	0.55 / 1.18	0.85 / 1.61	0.69 / 1.36	1.02 / 1.86	1.05 / 1.91	1.05 / 1.74	1.10 / 1.84	**1.17 / 2.08**
	1920x1080	ParkScene	0.14 / 0.38	0.14 / 0.50	0.46 / 1.23	0.59 / 1.47	0.49 / 1.21	0.64 / 1.58	0.70 / 1.69	0.67 / 1.64	0.69 / 1.70	**0.72 / 1.78**
		Cactus	0.19 / 0.38	0.20 / 0.48	0.50 / 1.00	0.77 / 1.38	0.62 / 1.15	0.83 / 1.49	0.85 / 1.52	0.83 / 1.42	0.82 / 1.48	**0.90 / 1.62**
		BQTerrance	0.20 / 0.28	0.20 / 0.38	0.40 / 0.67	0.63 / 1.06	0.50 / 0.87	0.65 / 1.06	0.72 / 1.23	0.70 / 1.18	0.72 / 1.14	**0.76 / 1.25**
		BasketballDrive	0.23 / 0.55	0.25 / 0.58	0.47 / 0.83	0.75 / 1.23	0.60 / 1.04	0.88 / 1.67	0.94 / 1.50	0.91 / 1.38	0.92 / 1.41	**1.07 / 1.65**
	Class C	RaceHorses	0.22 / 0.43	0.25 / 0.65	0.39 / 0.80	0.55 / 1.35	0.40 / 0.88	0.48 / 1.23	0.55 / 1.53	0.57 / 1.36	0.54 / 1.50	**0.69 / 1.80**
	832x480	BQMall	0.28 / 0.68	0.28 / 0.68	0.62 / 1.20	0.99 / 1.80	0.74 / 1.44	1.09 / 1.97	1.19 / 2.12	1.17 / 2.05	1.08 / 2.08	**1.28 / 2.28**
		PartScene	0.11 / 0.38	0.13 / 0.48	0.36 / 1.18	0.68 / 1.94	0.51 / 1.46	0.66 / 1.88	0.79 / 2.24	0.76 / 2.25	0.77 / **2.46**	**0.81** / 2.23
		BasketballDrill	0.25 / 0.58	0.33 / 0.68	0.58 / 1.20	0.79 / 1.49	0.66 / 1.27	0.88 / 1.67	0.99 / 1.89	0.95 / 1.79	0.99 / 1.78	**1.09 / 2.08**
	Class D	RaceHorses	0.27 / 0.55	0.31 / 0.73	0.59 / 1.43	0.83 / 2.08	0.60 / 1.44	0.85 / 2.11	0.95 / 2.44	0.94 / 2.19	0.95 / 2.35	**1.07 / 2.69**
	416x240	BQSquare	0.08 / 0.08	0.13 / 0.18	0.34 / 0.65	0.94 / 1.25	0.79 / 1.14	1.05 / 1.39	**1.28** / 1.72	1.08 / 1.45	1.14 / **1.81**	**1.28** / 1.70
		BlowingBubbles	0.16 / 0.35	0.18 / 0.58	0.53 / 1.70	0.74 / 2.26	0.62 / 1.95	0.78 / 2.40	0.86 / 2.67	0.82 / 2.58	0.84 / 2.43	**0.89 / 2.70**
		BasketballPass	0.26 / 0.58	0.31 / 0.75	0.73 / 1.55	1.08 / 2.12	0.85 / 1.75	1.12 / 2.23	1.26 / 2.51	1.22 / 2.32	1.14 / 2.31	**1.38 / 2.70**
	Class E	FourPeople	0.37 / 0.50	0.39 / 0.60	0.73 / 0.95	0.94 / 1.17	0.95 / 1.12	1.13 / 1.36	1.12 / 1.37	1.11 / 1.32	1.09 / 1.34	**1.17 / 1.43**
	1280x720	Johnny	0.25 / 0.10	0.32 / 0.40	0.60 / 0.68	0.81 / 0.88	0.75 / 0.85	0.90 / 0.94	0.89 / 0.98	0.90 / 0.97	0.86 / 0.96	**0.98 / 1.05**
		KristenAndSara	0.41 / 0.50	0.42 / 0.60	0.75 / 0.85	0.97 / 0.96	0.69 / 0.91	1.19 / 1.15	1.18 / 1.14	1.15 / 1.05	1.08 / 1.09	**1.21 / 1.20**
	Average		0.23 / 0.45	0.26 / 0.58	0.56 / 1.09	0.83 / 1.51	0.69 / 1.25	0.91 / 1.62	0.98 / 1.79	0.95 / 1.68	0.95 / 1.74	**1.06 / 1.89**
32	**Average**		0.18 / 0.19	0.26 / 0.35	0.51 / 0.68	0.86 / 1.04	0.67 / 0.83	0.87 / 1.07	0.99 / 1.24	0.98 / 1.20	0.96 / 1.22	**1.12 / 1.40**
27	**Average**		0.16 / 0.09	0.33 / 0.26	0.49 / 0.42	0.72 / 0.57	0.63 / 0.52	0.82 / 0.68	0.97 / 0.81	0.97 / 0.80	0.87 / 0.73	**1.10 / 0.90**
22	**Average**		0.13 / 0.04	0.27 / 0.14	0.46 / 0.27	0.63 / 0.34	0.55 / 0.29	0.76 / 0.42	0.87 / 0.48	0.85 / 0.48	0.84 / 0.47	**0.95 / 0.54**

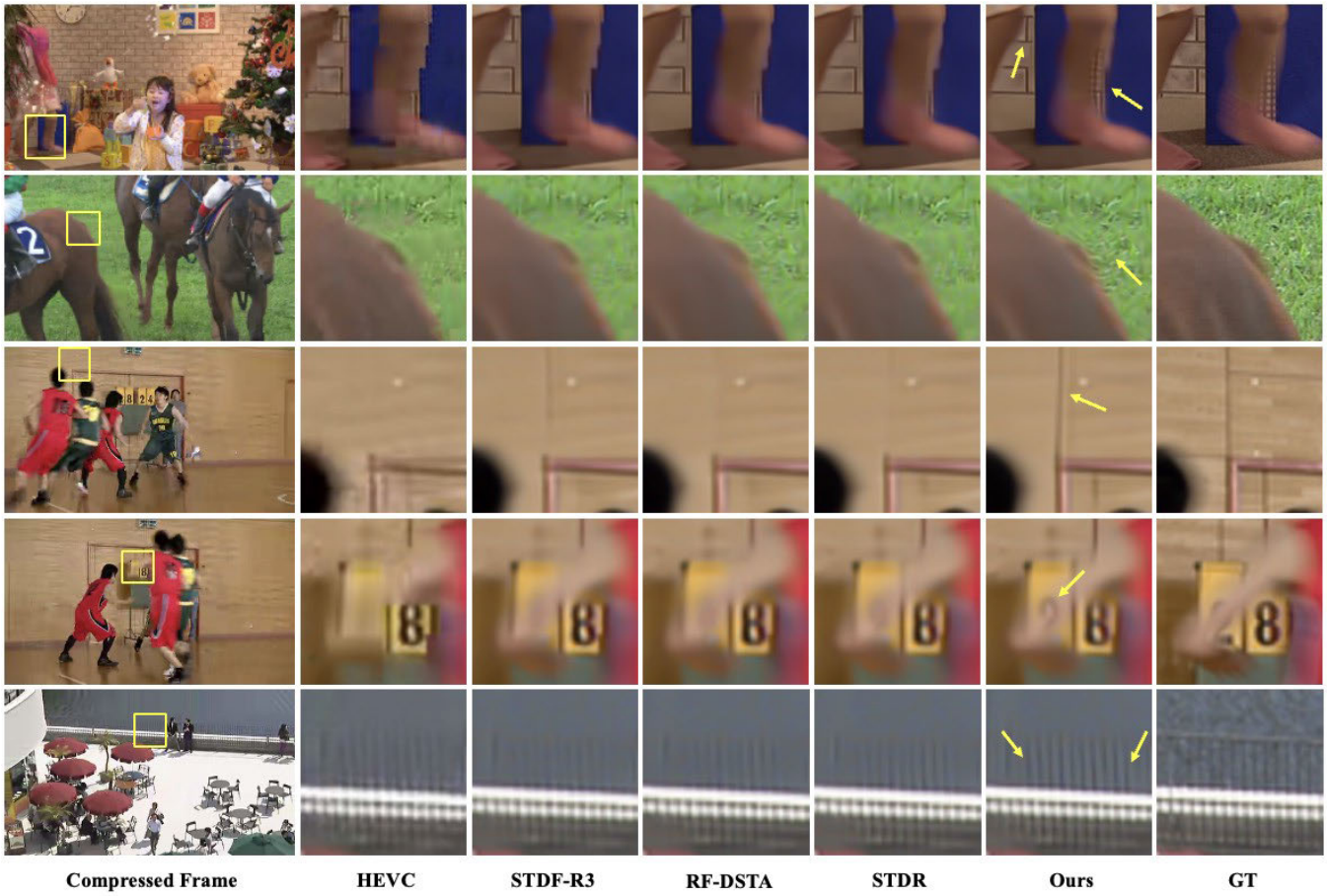


Fig. 7. Subjective comparison over four standard test sequences at QP = 37. Compared with existing methods, boundary of fast-moving object become clearer (the fourth row). At the same time, structural information with high frequency components is enhanced (the third row). Furthermore, textured regions with rich non-high frequency features are also enhanced significantly (the first, second and last row). Zoom in for best view.

Fig. 7 provides the qualitative performance on four testing sequences with large displacement motion and artifacts with different frequencies. Compared with compressed sequences [1] and existing methods (STDF-3L [10], RF-

TABLE II

OVERALL PERFORMANCE COMPARISON FOR MODEL SIZE (M), MACS (G) AT RESOLUTION 416×240, $\triangle$PSNR (dB) AND $\triangle$SSIM ($\times 10^{-2}$) OVER STANDARD TEST SEQUENCES AT QP = 37

Method	Size	MACs	△PSNR	△SSIM
STDF*	4.71	304.06	0.90	1.65
BasicVSR++*	3.45	1647.77	0.97	1.79
OVQE*	3.11	1617.02	1.01	1.83
Ours	3.55	1143.37	**1.06**	**1.89**

TABLE III

OVERALL PERFORMANCE COMPARISON FOR $\triangle$PSNR (dB) AND $\triangle$SSIM ($\times 10^{-2}$) OVER STANDARD TEST SEQUENCES AT QP = 37. INTRA-CLASS AVERAGE VALUES ARE CALCULATED FOR FIVE DIFFERENT RESOLUTIONS (A-E). THE AVERAGE VALUES OF ALL TEST SEQUENCES ARE SHOWN IN THE LAST ROW

Sequence	STDF*	BasicVSR++*	OVQE*	Ours
Class A	1.08 / 1.72	1.16 / 1.82	1.19 / 1.84	**1.27 / 1.93**
Class B	0.81 / 1.51	0.86 / 1.57	0.89 / 1.62	**0.92 / 1.68**
Class C	0.79 / 1.72	0.86 / 1.95	0.90 / 1.99	**0.97 / 2.10**
Class D	0.96 / 2.08	1.05 / 2.32	1.10 / 2.38	**1.16 / 2.45**
Class E	1.02 / 1.16	1.07 / 1.19	1.09 / 1.20	**1.12 / 1.23**
Total	0.90 / 1.65	0.97 / 1.79	1.01 / 1.83	**1.06 / 1.89**

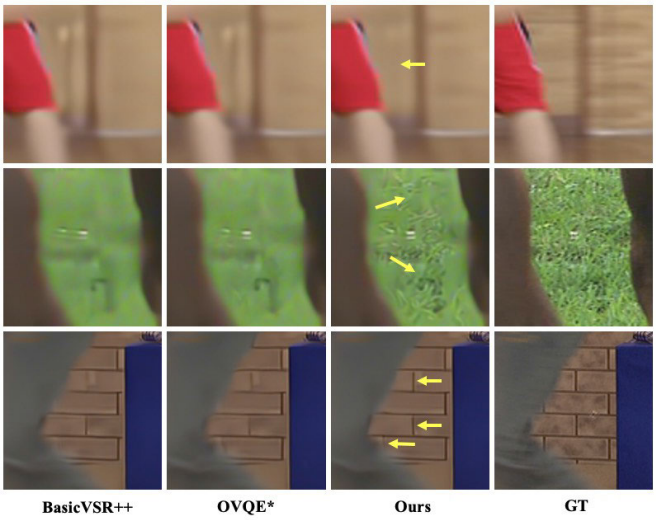


Fig. 8. Subjective comparison over three standard test sequences at QP = 37. Compared with BasicVSR++* and OVQE*, structural information with high frequency components is enhanced (the first and last row). Moreover, textured regions with rich non-high frequency features are also enhanced significantly (the second row). Zoom in for best view.

DSTA [23], and STDR [25]), blurred motion boundaries become clearer in our results. As shown in the first row of the figure, the texture information reconstruction of the box illustrates that our method has significant advantages in inter-frame feature alignment. The results in the second and last rows illustrate that our method is able to enhance features of different frequencies, such as grass and railings with rich textures. The third row shows the enhanced results of the high-frequency structural information. In particular, the results in the fourth row show that our method can alleviate motion blur, even though the raw frame itself has motion blur. Removing the motion blur of the raw video itself is beneficial for improving the subjective experience enhancement of the compressed video, which has not been discussed by existing work. More subjective results are detailed in Fig. 8.

2) Frequency Enhancement: Regular convolutional neural networks tend to focus on high-frequency features with noticeable gradient changes. The learned filter operator exhibits a specific and singular direction in the spatial domain, thus ignoring non-high-frequency features in other directions. The above defects limit the performance of existing methods, which are analyzed in detail in Fig. 2. In our proposed MDC module, the sampling position of the convolution kernel is fixed skillfully in the spatial domain. This design can respond to the feature with different frequencies effectively.

To verify the effectiveness more intuitively, we select some test sequences with rich non-high frequency information to conduct both objective and subjective experiments. As shown in Fig. 11, it shows that both the PSNR and SSIM of our method are higher than those of OVQE* on these patches. The first row represents patches with both high-frequency information (arrows) and non-high-frequency information (dashed boxes). Compared with the quality improvement of high-frequency features in existing methods, our method can enhance non-high-frequency information. The second row shows the part after separating the non-high frequency area. Textured regions with non-high frequencies become clearer in our method.

3) Rate-Distortion: The Rate-distortion (RD) curve is usually used to evaluate the performance of video codec standards or algorithms. It reflects the relationship between video quality and the number of bits occupied by compressed video per second during transmission. In other words, it can also be regarded as a tool for evaluating the performance of compressed video quality enhancement algorithms. In the case of the *BasketballPass* sequence, we collect the average PSNR value from the compression log file with four QPs, i.e., 22, 27, 32, and 37. As shown in Fig. 9, we can see that the PSNR value of our method is higher than that of other methods (HEVC, STDF-R3, and STDR) under the same code stream overhead.

4) Quality Fluctuation: Quality fluctuation, which reflects the range of quality variation between frames, is another commonly used objective criterion for video quality evaluation. Generally, slight fluctuation indicate the coherence and stability of the video. Quality fluctuations are evaluated by calculating the standard deviation (SD) of the PSNR values. Here, we select a sequence with long-range inter-frame motion *BasketballPass* as an example. Fig. 12 shows that sthe PSNR values of a clip with 40 frames is plotted as a line graph. The SD values of HEVC, STDF-R3, STDR, and our method are 1.77, 0.99, 0.90, and 0.88, respectively. The experimental results show that our method can further alleviate the quality fluctuations of compressed video.

5) Performance on Hybrid-Distortion: To evaluate the enhancement ability of our method on other hybrid-distorted video, we select several interesting video sequences with different resolution from the Bilibili video website and YouTube website. The sequences are corroded by multiple types of distortion, including motion blur, jagged edges, and compression artifacts. We present the enhancement results for the visual experience due to the lack of high-quality raw

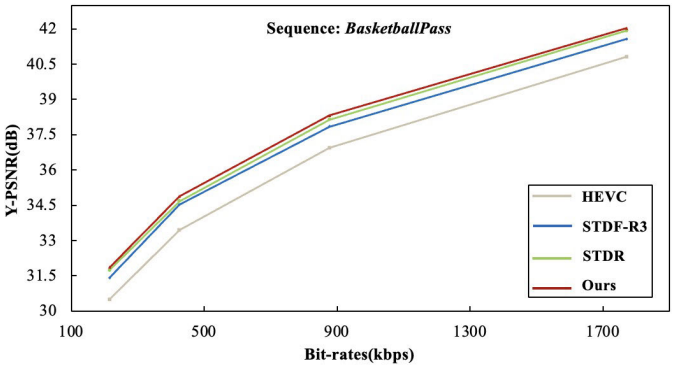


Fig. 9. The Rate-distortion (RD) curve between video quality and the number of bits occupied by compressed video per second. Zoom in for best view.

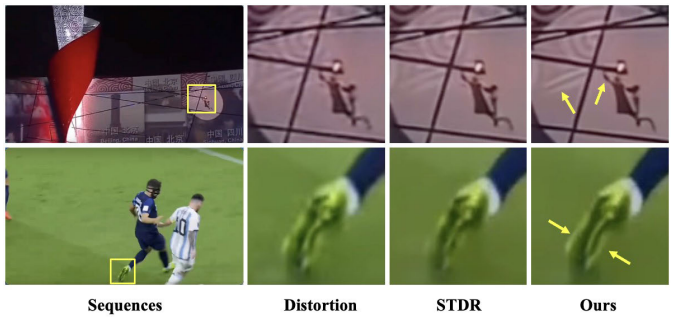


Fig. 10. Subjective comparison of enhancement results for low-quality video sequences with hybrid-distortion. Compared with STDR, both the texture information and the boundary of moving object become clearer and sharper in our method. Furthermore, the noise in the video is well removed. Zoom in for best view.

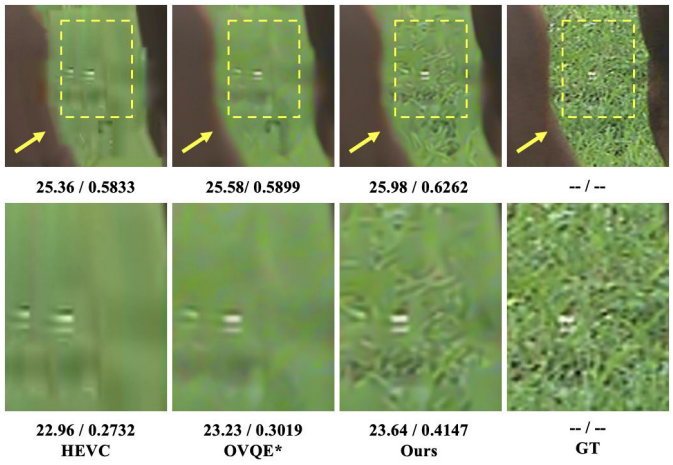


Fig. 11. Frequency Enhancement. The first column represent the non-high-frequency patch clipped from compressed frames under $QP = 37$. The second and third columns represent the enhanced results by the method OVQE* and ours, respectively. The raw patch is shown in the last column. PSNR (dB) / SSIM are chosen as objective quality evaluation metric.

video. Fig. 10 shows that our method has a better subjective visual experience in both the texture and structure regions than that of STDR. In particular, our model effectively mitigates compression artifacts, such as blockiness and motion blur.

D. Ablation Experiment

To verify the performance of the proposed three modules MSADA, PIPA, and MDFE, the baselines are set as follows:

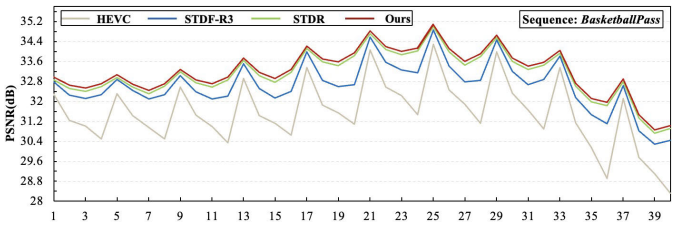


Fig. 12. Quality fluctuation. Fluctuation trend of PSNR values for the sequence *BasketballPass*. Zoom in for best view.

TABLE IV

OVERALL PERFORMANCE COMPARISON FOR △PSNR (dB) AND △SSIM ($\times 10^{-2}$) OVER STANDARD TEST SEQUENCES AT QP = 37

MSADA	PIPA	MDFE	Size	△PSNR	△SSIM
–	–	–	0.99	0.79	1.47
√	–	–	1.13	0.85	1.53
–	√	–	1.39	0.84	1.55
–	–	√	2.74	0.93	1.71
–	√	√	3.42	1.01	1.80
√	–	√	3.67	1.02	1.78
√	√	–	2.81	0.98	1.72
√	√	√	3.55	**1.06**	**1.89**

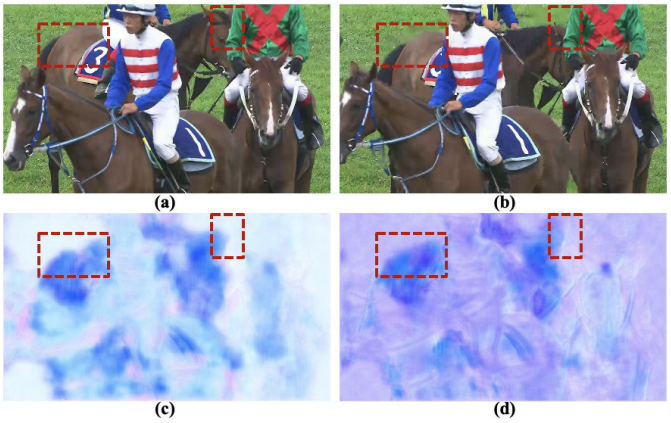


Fig. 13. Visualization of offset predicted by the proposed **MSADA** module. (a)Reference frame. (b)Target frame. (c)Offset of regular deformable convolution. (d) Offset of atrous deformable convolution.

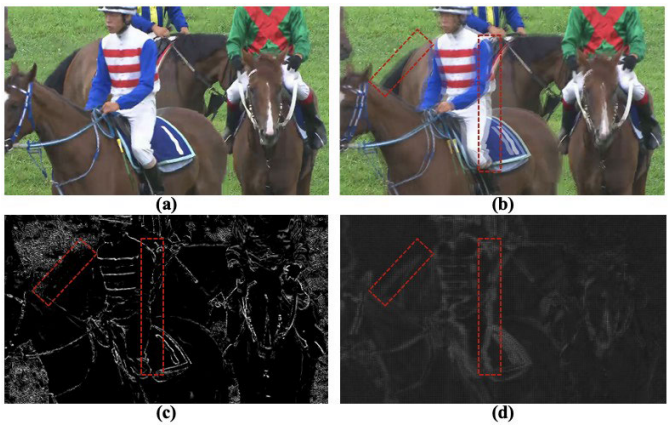


Fig. 14. Visualization of feature map produced by the proposed **PIPA** module. (a)Reference frame. (b)Overlay of target frame and reference frame. (c)Feature map of pixel-wise deformable convolution. (d) Feature map of patch-wise deformable convolution.

regular deformable kernel without the MDFE module in the STFA unit, a pixel-wise deformable convolution without the MDFE module in the BDFP unit and a residual block without the MDFE module in the RFE unit. We also adjust the number

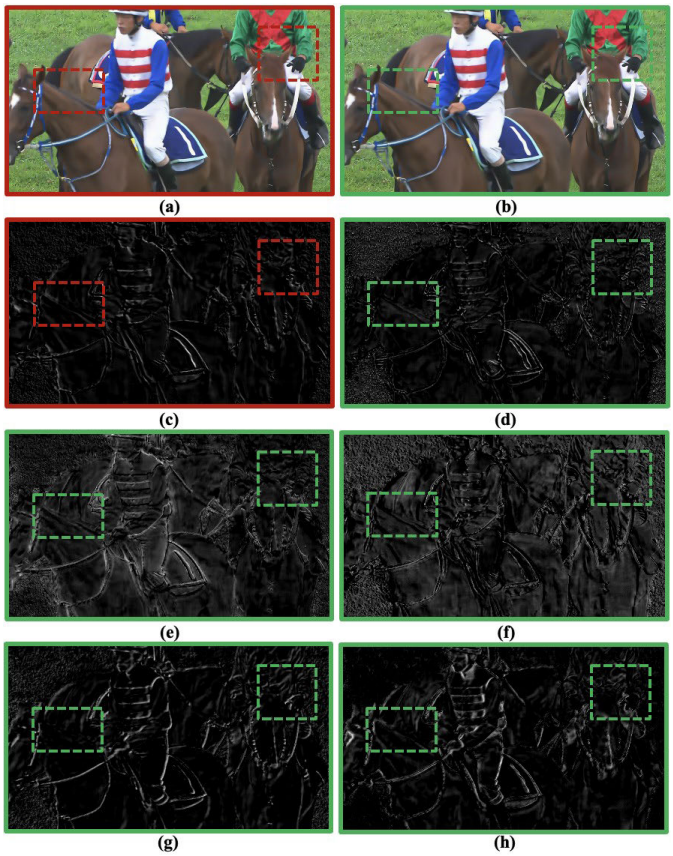


Fig. 15. Visualization of feature maps produced by the proposed **MDC** module. (a) Enhanced result of STDR. (b) Enhanced result of ours. (c) Feature maps of STDR. (d-h) Feature maps of ours. Compared with STDR, our method can capture sufficient features with spatial diversity and frequency diversity. Zoom in for best view.

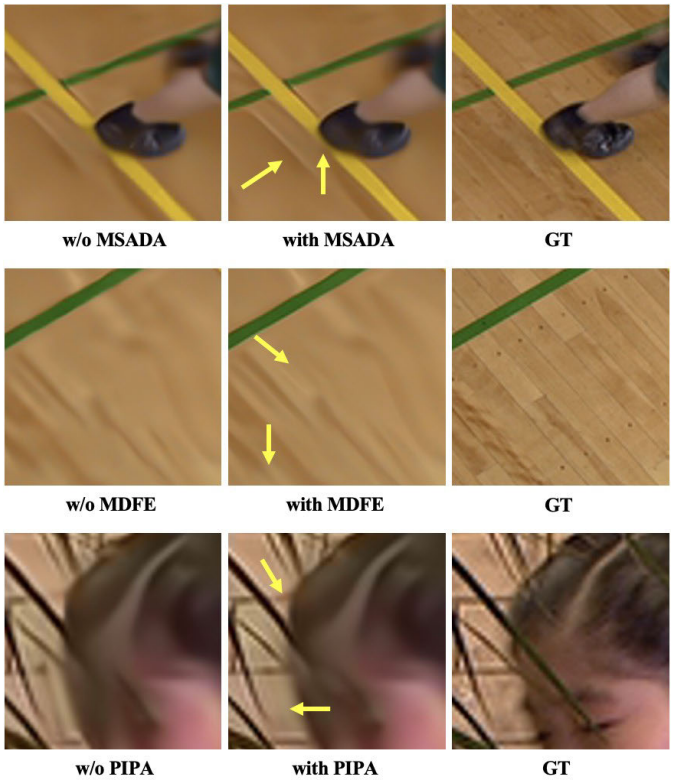


Fig. 16. Effectiveness of the proposed modules: **MSADA**, **PIPA** and **MDFE**. The first row in the figure indicates that the **MSADA** module is able to capture non-local motion information. The structure near the shoe has been further improved. The second row in the figure shows that the **MDFE** module can enhance features with different frequency. Mid and low frequency textures in ground areas become clearer. The last row in the figure proves that the **PIPA** module can obtain spatio-temporal information with better non-local consistency. The ringing effect on the girl's face is effectively removed.

of feature channels to achieve a similar parameter size for a fair comparison. The overall experimental results are listed in Table. IV.

1) Effectiveness of MSADA: The comparative experiments with and without the MSADA module are listed in the first row, second row, fifth row and last row of the Table. IV. The experimental results show that the introduced module can further improve the quality of compressed frames. We visualize the corresponding predicted offset in Fig. 13 to intuitively illustrate how atrous deformable convolution with multiple dilation factors works effectively. Compared with the offset in regular convolution, the estimated offset in our method can better take into account the boundary of moving objects. This process accounts for quality improvement. The subjective results are presented in Fig. 16.

2) Effectiveness of PIPA: The comparative experiments with and without the PIPA module are listed in the first row, third row, sixth row and last row of the Table. IV. The experimental results show that the introduced module can further improve the quality of compressed frames. We visualize its extracted features in Fig. 14 to intuitively illustrate the effectiveness of patchwise deformable alignment. Compared with the feature in pixel-wise deformable convolution, the feature in our method can describe the non-local feature alignment of moving objects. Combined with non-local and local spatio-temporal information, our method can effectively remove motion artifacts caused by compression. The subjective results are presented in Fig. 16.

3) Effectiveness of MDFE: The results of comparative experiments with and without the MDFE module are listed in the first row, fourth row, seventh row and last row of the Table. IV. This shows that the MDFE module dramatically improves the quality of compressed frames. We visualize its extracted features in Fig. 15 to intuitively illustrate the effectiveness of MDC. For fairness, we remove the features under the first feature fusion layer after feature alignment and calculate the average of channel features. For the horse neck region in the boxes, the features provided by our method are more accurate and richer. With the help of MDC, our method can capture sufficient features with spatial diversity and frequency diversity. The subjective results are presented in Fig. 16.

In addition to conducting ablation experiments on the overall performance of each proposed module, we consider the impact of hyperparameters or setting choices within the module. Whether the existing default settings in our method are optimal will be discussed. Additionally, the performance fluctuations that may be caused by adjusting the existing setting are analyzed further.

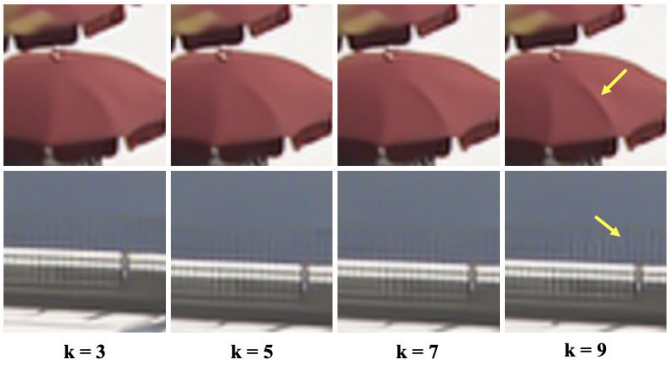


Fig. 17. Subjective results of ablation experiments on kernel size in MDC. Both high-frequency components (the first row) and non-high-frequency components (the last row) are further enhanced when $k = 9$.

TABLE V

ABLATION EXPERIMENT OF KERNEL SIZE SETTING IN MDC. IT INCLUDES MODEL SIZE (M), MACS (G) AT RESOLUTION 416×240, △PSNR (dB) AND △SSIM ($\times 10^{-2}$) OVER STANDARD TEST SEQUENCES AT QP = 37

Setting	△PSNR / △SSIM	Size	MACs
k=3	1.06 / 1.89	3.55 (2.60)	1143.37
k=5	1.06 / 1.93	5.22 (2.70)	1212.07
k=7	1.07 / 1.91	7.68 (2.80)	1280.77
k=9	1.10 / 2.02	10.92 (2.89)	1349.48

TABLE VI

ABLATION EXPERIMENT FOR DIRECTION SETTING IN MDC. IT INCLUDES MODEL SIZE (M), MACS (G) AT RESOLUTION 416×240, △PSNR (dB) AND △SSIM ($\times 10^{-2}$) OVER STANDARD TEST SEQUENCES AT QP = 37

Setting	△PSNR / △SSIM	Size	MACs
(a), (b), (c), (d), (e)	1.06 / 1.89	3.55 (2.60)	1143.37
(a), (a), (a), (a), (a)	1.00 / 1.78	3.40 (3.34)	1658.64
(a)-(e), (f), (g)	1.05 / 1.91	4.61 (2.77)	1212.61
(a)-(k)	1.06 / 1.94	6.80 (3.19)	1351.09

TABLE VII

ABLATION EXPERIMENT OF DILATION FACTOR IN MSADA. IT INCLUDES MODEL SIZE (M), MACS (G) AT RESOLUTION 416×240, △PSNR (dB) AND △SSIM ($\times 10^{-2}$) OVER STANDARD TEST SEQUENCES AT QP = 37

Setting	△PSNR / △SSIM	Size	MACs
d = 1, 3, 5	1.06 / 1.89	3.55 (2.60)	1143.37
d = 1, 1, 1	1.01 / 1.82	3.55 (2.60)	1143.37
d = 1, 3, 5, 7	1.07 / 1.96	3.62 (2.66)	1184.32
d = 1, 3, 5, 7, 9	1.05 / 1.92	3.68 (2.72)	1225.27

4) Kernel Size in MDC: We propose the MDFE module to respond to features with different frequencies simultaneously to alleviate the phenomenon in which existing methods focus on the enhancement of high-frequency information. Specifically, we achieve this goal by including a convolution layer in five different directions, which are shown in the subfigures (a-e) in Fig. 5. The default value of the convolution kernel size k in our method is 3.

In fact, the choice of kernel size plays a vital role in improving the performance of CNNs. Combined with the MDC module, we adjust the kernel size to 5, 7, and 9 in sequence. The corresponding ablation experiment results are shown in Table. V. Model size and multiple accumulate operations (MACs) are also calculated for complexity comparison. The values in brackets in the third column represent the number of parameters actually trained and updated in our model. For example, the position parameters represented by "0" in Fig. 5 are masked out after initialization. In other words, the number of parameters actually updated in our method is only 2.6M, which can be illustrated by the comparison of the MACs in Table. II.

The results in the fourth row show that the performance is significantly improved when $k = 9$. We speculate that the convolution receptive field under this setting breaks through the minimum value of the block partition unit (8×8) in the video compression algorithm. Larger receptive fields in different directions enable the model to learn inter-block motion information efficiently. The above experimental results also verify the novelty of our module design based on the characteristics of video compression. Benefiting from the proposed MDC module, the actual number of updated parameter slightly increases when $k = 9$, which can also be explained by the computational overhead in the last column. Moreover, the subjective results for adjusting the kernel size are shown in Fig. 17.

5) Direction Setting in MDC: In addition to the selection of the convolution kernel size in the proposed MDC module, it is worth discussing whether adding different kinds of direction convolutions can further improve the performance. As shown in Fig. 5, we design six other types of directions convolution (f-k), which are derived from the combinations of the five existing convolution types.

In Table. VI, the experimental results of three sets of settings in the multi-direction convolution (MDC) module are listed. Model size and MACs are also calculated for complexity comparison. The results of the first two rows show that multi-direction convolution can bring higher quality improvements compared to regular convolution kernels. The results in the first, third, and fourth rows show that adding multi-direction convolution types does not significantly improve the performance. We believe that the expanded direction convolution kernel (f-k) comprises the existing five basic types (a-e). Moreover, further expansion of the direction convolution will gradually make it homogeneous with the regular convolution kernel. Due to computational complexity, the existing MDC module setting is considered the optimal choice.

6) Dilation Factor in MSADA: Our proposed MSADA module adopts atrous deformable convolution instead of regular deformable convolution. Deformable convolution with a dilation factor has advantages in obtaining spatio-temporal information over a broader range. Is there an upper limit to the performance gain caused by a larger dilation factor? We verify this through ablation experiments on the dilation factor. Compared with the default dilation factor $k = 1, 3, 5$, we also consider the setting $k = 7, 9$.

In Table. VII, the second row represents the results of regular deformable convolution ($d = 1$) with the same model size and MACs. The experimental results in the first two rows show that the larger range of atrous deformable convolution in our method can further improve the quality improvement.

However, the quality metric begins to decrease as the dilation factor increases. We suspect that when the distance between the sampling point and the center point increases, the training of the deformable convolution will becomes unstable. As a first step in our approach, the MSADA module receives compressed frames with severe blocking artifacts, limiting the performance gains from a wider range of sampling points. Therefore, we still adopt the existing dilation factor $k = 1, 3, 5$.

V. Conclusion

This paper proposes an enlarged motion-aware and frequency-aware network (EMAFA) for compressed video artifact reduction. The proposed framework contains three novel key modules: the multi-scale atrous deformable alignment (MSADA) module, the pixel-wise and patch-wise deformable alignment (PIPA) module, and the multi-direction frequency enhancement (MDFE) module. To obtain local and global motion receptive fields, the MSADA module with enlarged flexible sampling positions and the PIPA module are designed to perceive spatio-temporal information from long-range displacement motion. Moreover, the plug and play MDFE module with multiple direction convolution can improve quality consistency across regions with different frequencies. Extensive experiments show that the proposed method achieves better performance than other state-of-the-art methods. Supplementary experiments also demonstrate the considerable generalizability of the proposed method for improving hybrid-distortion quality.

References

[1] 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.

[2] X. Meng, X. Deng, S. Zhu, X. Zhang, and B. Zeng, "A robust quality enhancement method based on joint spatial–temporal priors for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2401–2414, Jun. 2021.

[3] Y. Wang, X. Fan, S. Liu, D. Zhao, and W. Gao, "Multi-scale convolutional neural network-based intra prediction for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1803–1815, Jul. 2020.

[4] R. Yang, H. Liu, S. Zhu, X. Zheng, and B. Zeng, "DFCE: Decoder-friendly chrominance enhancement for HEVC intra coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1481–1486, Mar. 2023.

[5] 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.

[6] 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.

[7] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *Int. J. Comput. Vis.*, vol. 127, no. 8, pp. 1106–1125, Aug. 2019.

[8] 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.

[9] D. Ding, W. Wang, J. Tong, X. Gao, Z. Liu, and Y. Fang, "Biprediction-based video quality enhancement via learning," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1207–1220, Feb. 2022.

[10] 003 J. Deng, L. Wang, S. Pu, and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10696–10703.

[11] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "EDVR: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 1954–1963.

[12] T. Liu, M. Xu, S. Li, R. Ding, and H. Liu, "MRS-Net+ for enhancing face quality of compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2881–2894, May 2022.

[13] H. Zhao et al., "CBREN: Convolutional neural networks for constant bit rate video quality enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4138–4149, Jul. 2022.

[14] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Trans. circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 614–619, Jul. 2003.

[15] A.-C. Liew and H. Yan, "Blocking artifacts suppression in block-coded images using overcomplete wavelet representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 4, pp. 450–461, Apr. 2004.

[16] R. Yang, M. Xu, T. Liu, Z. Wang, and Z. Guan, "Enhancing quality for HEVC compressed videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 2039–2054, Jul. 2019.

[17] 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.

[18] Z. Jin, P. An, C. Yang, and L. Shen, "Quality enhancement for intra frame coding via CNNs: An adversarial approach," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 1368–1372.

[19] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[20] Y. Xu, L. Gao, K. Tian, S. Zhou, and H. Sun, "Non-local convLSTM for video compression artifact reduction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct./Nov. 2019, pp. 7043–7052.

[21] Q. Ding, L. Shen, L. Yu, H. Yang, and M. Xu, "Patch-wise spatial–temporal quality enhancement for HEVC compressed video," *IEEE Trans. Image Process.*, vol. 30, pp. 6459–6472, 2021.

[22] W. Bao, W.-S. Lai, X. Zhang, Z. Gao, and M.-H. Yang, "MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 933–948, Mar. 2021.

[23] M. Zhao, Y. Xu, and S. Zhou, "Recursive fusion and deformable spatiotemporal attention for video compression artifact reduction," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5646–5654.

[24] L. Peng, A. Hamdulla, M. Ye, S. Li, Z. Wang, and X. Li, "OVQE: Omniscient network for compressed video quality enhancement," *IEEE Trans. Broadcast.*, vol. 69, no. 1, pp. 153–164, Mar. 2023.

[25] D. Luo, M. Ye, S. Li, C. Zhu, and X. Li, "Spatio-temporal detail information retrieval for compressed video quality enhancement," *IEEE Trans. Multimedia*, vol. 25, pp. 6808–6820, Oct. 2022.

[26] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12179–12188.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[28] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[29] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable ConvNets V2: More deformable, better results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 9308–9316.

[30] H. Wang, X. Xiang, Y. Tian, W. Yang, and Q. Liao, "STDAN: Deformable attention network for space-time video super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–11, Feb. 2023.

[31] S. Xu, B. Song, X. Chen, and J. Zhou, "Direction-aware video demoiréing with temporal-guided bilateral learning," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 6, pp. 6360–6368.

[32] H. Zhang, H. Xie, and H. Yao, "Spatio-temporal deformable attention network for video deblurring," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Berlin, Germany: Springer, Oct. 2022, pp. 581–596.

[33] B. Jiang, Z. Xie, Z. Xia, S. Li, and S. Liu, "ERDN: Equivalent receptive field deformable network for video deblurring," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, Tel Aviv, Israel. Berlin, Germany: Springer, Oct. 2022, pp. 663–678.

[34] 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.

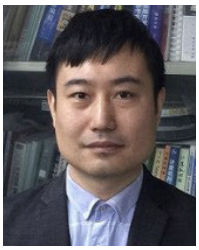
[35] Z. Wang, M. Ye, S. Li, and X. Li, "Multi-frame compressed video quality enhancement by spatio-temporal information balance," *IEEE Signal Process. Lett.*, vol. 30, pp. 105–109, 2023.
[36] K. Wang, F. Chen, Z. Ye, L. Wang, X. Wu, and S. Pu, "A spatio-temporal decomposition network for compressed video quality enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.
[37] L. Wang, Y. Guo, L. Liu, Z. Lin, X. Deng, and W. An, "Deep video super-resolution using HR optical flow estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 4323–4336, 2020.
[38] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4472–4480.
[39] X. Ying, L. Wang, Y. Wang, W. Sheng, W. An, and Y. Guo, "Deformable 3D convolution for video super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 1500–1504, 2020.
[40] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[41] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.
[42] S. Liu, C. Li, N. Nan, Z. Zong, and R. Song, "MMDM: Multi-frame and multi-scale for image demoireing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 434–435.
[43] K. C. K. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving video super-resolution with enhanced propagation and alignment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5972–5981.
[44] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud, "Two deterministic half-quadratic regularization algorithms for computed imaging," in *Proc. 1st Int. Conf. Image Process.*, vol. 2, 1994, pp. 168–172.
[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

Wang Liu is currently pursuing the Ph.D. degree with the School of Electronic and Computer Engineering, Peking University. His research interests include video/point cloud enhancement and processing.

Wei Gao (Senior Member, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, in February 2017. In 2016, he was a Visiting Scholar with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. From 2017 to 2019, he was a Post-Doctoral Fellow with the City University of Hong Kong and a Research Fellow with Nanyang Technological University, Singapore. Since 2019, he has been an Assistant Professor with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, Shenzhen, China. His research interests include point cloud compression and processing, image/video coding and processing, deep learning, and artificial intelligence. He is currently an Elected Member of the Visual Signal Processing and Communications Technical Committee, the IEEE Circuits and Systems Society, the Image Video and Multimedia Technical Committee, and the Asia-Pacific Signal and Information Processing Association. He is serving as an Associate Editor for *Signal Processing* (Elsevier).

Ge Li (Member, IEEE) received the B.E. degree from the Department of Computer Science and Engineering, Dalian University of Technology, Dalian, China, in 1988, and the M.S. degree from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 1991. From 1991 to 1992, he was a Lecturer with the Department of Electrical Engineering, Dalian Maritime University, Dalian. He was a Summer Engineer with CSG Research Laboratories, Motorola Inc., Libertyville, IL, USA, in 1995. He is currently a Professor with the School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China. His research interests include image/video process and analysis, machine learning, digital communications, and signal processing.

Siwei Ma (Fellow, IEEE) received the B.S. degree from Shandong Normal University, Jinan, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2005. From 2005 to 2007, he held a post-doctoral position with the University of Southern California, Los Angeles, CA, USA. He joined the School of Electronics Engineering and Computer Science, Institute of Digital Media, Peking University, Beijing, where he is currently a Professor. He has authored more than 200 technical articles in refereed journals and proceedings in image and video coding, video processing, video streaming, and transmission. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *Journal of Visual Communication and Image Representation*.



Tiesong Zhao (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2006, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2011. He was a Research Associate with the Department of Computer Science, City University of Hong Kong, from 2011 to 2012; a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, from 2012 to 2013; and a Research Scientist with the Ubiquitous Multimedia Laboratory, State University of New York at Buffalo, from 2014 to 2015. He is currently a Minjiang Distinguished Professor with the College of Physics and Information Engineering, Fuzhou University, China. His research interests include multimedia signal processing, coding, quality assessment, and transmission. Due to his contributions in video coding and transmission, he received Fujian Science and Technology Award for Young Scholars in 2017. He has been serving as an Associate Editor for *IET Electronics Letters* since 2019.

Hui Yuan (Senior Member, IEEE) received the B.E. and Ph.D. degrees in telecommunication engineering from Xidian University, Xi'an, China, in 2006 and 2011, respectively. From January 2013 to December 2014, he was a Post-Doctoral Fellow with the Department of Computer Science, City University of Hong Kong. Since 2014, he has been a Lecturer, an Associate Professor, and a Full Professor with Shandong University, Jinan, China. His research interests include video/image/immersive media compression, adaptive video streaming, and computer vision.