Biprediction-Based Video Quality Enhancement via Learning

Dandan Ding, Wenyu Wang, Junchao Tong, Xinbo Gao, *Senior Member, IEEE*, Zoe Liu, and Yong Fang

***Abstract*—Convolutional neural networks (CNNs)-based video quality enhancement generally employs optical flow for pixelwise motion estimation and compensation, followed by utilizing motion-compensated frames and jointly exploring the spatiotemporal correlation across frames to facilitate the enhancement. This method, called the optical-flow-based method (OPT), usually achieves high accuracy at the expense of high computational complexity. In this article, we develop a new framework, referred to as biprediction-based multiframe video enhancement (PMVE), to achieve a one-pass enhancement procedure. PMVE designs two networks, that is, the prediction network (Pred-net) and the frame-fusion network (FF-net), to implement the two steps of synthesization and fusion, respectively. Specifically, the Pred-net leverages frame pairs to synthesize the so-called virtual frames (VFs) for those low-quality frames (LFs) through biprediction. Afterward, the slowly fused FF-net takes the VFs as the input to extract the correlation across the VFs and the related LFs, to obtain an enhanced version of those LFs. Such a framework allows PMVE to leverage the cross-correlation between successive frames for enhancement, hence capable of achieving high accuracy performance. Meanwhile, PMVE effectively avoids the explicit operations of motion estimation and compensation, hence greatly reducing the complexity compared to OPT. The experimental results demonstrate that the peak signal-to-noise ratio (PSNR) performance of PMVE is fully on par with that of OPT while its computational complexity is only 1% of OPT. Compared with other state-of-the-art methods in the literature, PMVE is also confirmed to achieve superior performance in both objective quality and visual quality at a reasonable complexity level. For instance, PMVE can surpass its best counterpart method by up to 0.42 dB in PSNR.**

***Index Terms*—Convolutional neural network (CNN), frame prediction, multiframe, optical flow, video enhancement.**

Manuscript received October 16, 2019; revised March 4, 2020; accepted May 18, 2020. Date of publication June 17, 2020; date of current version February 16, 2022. This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY20F010013, and in part by the Open Project of Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking. This article was recommended by Associate Editor Q. M. J. Wu. *(Corresponding author: Yong Fang.)*

Dandan Ding is with the School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China, and also with the Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking, Hangzhou 310027, China (e-mail: dandanding@hznu.edu.cn).

Wenyu Wang and Junchao Tong are with the School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China.

Xinbo Gao is with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (e-mail: gaoxb@cqupt.edu.cn).

Zoe Liu is with the Department of Research and Development, Visionular Inc., Mountain View, CA 94040 USA (e-mail: zoeliu@visionular.com).

Yong Fang is with the School of Information Engineering, Chang'an University, Xi'an 710064, China (e-mail: fy@chd.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2020.2998481.

Digital Object Identifier 10.1109/TCYB.2020.2998481

I. Introduction

In recent years, the demand for ultrahigh-definition (UHD) video is constantly on the rise. Video compression plays a crucial role in the storage and transmission of UHD videos. Generally, the encoder applies compression techniques to encode the original video into a compressed bitstream for bit-rate saving, and this bitstream is decoded by the decoder to reconstruct the video. The lossy compression, which is universally deployed today, introduces compression artifacts and inevitably causes quality degradation to the reconstructed video. An effective solution to reduce artifacts and improve the quality of the compressed video is through the use of the postprocessing scheme, applied to those decoded frames to further improve the video quality subjectively or objectively.

Most recently, the convolutional neural network (CNN) showed outstanding performance and led the trend in the exploration of image/video postprocessing [1]. CNN-based compressed video enhancement was initially inspired by super resolution (SR), which targets generating a high-resolution (HR) image from a low-resolution (LR) version. Various SR networks have been proposed [2]–[9]. By learning the nonlinear mapping between LR and HR images, these networks are then able to reconstruct the HR image from its LR counterpart. The image/video enhancement problem can be resolved in a similar way as done in SR. Following this spirit, a significant amount of CNN-based schemes have been proposed.

Dong *et al.* [10] first developed a four-layer artifacts reduction CNN (ARCNN) for artifact reduction over JPEG images. Later on, various kinds of advanced networks such as [11]–[15] are proposed. These networks are all originally designed for image restoration, hence unsuitable for compressed video enhancement because the coding techniques adopted in video compression are generally quite different from those for images. Recently, some special networks have been customized for video applications and achieved significant gains [16]–[33]. Details of this category of related work will be further addressed in the next section. These networks mainly adopt single-frame methods to enhance one frame by exploring the spatial correlation among pixels within the same frame. As opposed to the single-frame method, multiframe enhancement introduces a third dimension, that is, the temporal dimension, in addition to the two spatial dimensions that represent an image. This third dimension contributes a fairly large difference for all video-oriented methodologies compared to those originally designed for images. Along the temporal direction a high degree of correlation is inherited, which can be well-taken advantage of to help improve the performance of enhancement [34].

As such, the multiframe enhancement is challenging because it considers both spatial and temporal dimension information. It is critical to solve the following problems: 1) How to explore the temporal correlation across frames? and 2) How to design a network involving both spatial and temporal information? Another big challenge is the extra computational complexity introduced while leveraging the temporal information. A conventional solution is the optical flow-based method which consists of two steps: 1) frame alignment and 2) frame fusion. In frame alignment, pixelwise motion vectors between frames are estimated in order to align the neighboring frames to the current frame. In frame fusion, the aligned frames and the current frame are fused to generate an enhanced frame. Lu *et al.* [35] and Tong *et al.* [36] followed this manner to enhance the compressed videos. To boost the peak signal-to-noise ratio (PSNR) or structural similarity (SSIM) performance, they both adopt FlowNet [37] to acquire accurate optical flow, which greatly increases the computational complexity. Yang *et al.* [38] and Guan *et al.* [39] developed the multiframe quality enhancement (MFQE) to utilize the high-quality frames (HFs) neighboring the current low-quality frame (LF) to enhance the quality of the current frame. To reduce the computational burden, they employ lightweight spatial transformer motion compensation (STMC) [34] to generate the optical flow. Results show that the computational complexity is reduced but the quality improvement is also compromised. Bao *et al.* [40] adapted their motion estimation and motion compensation-driven neural network (MEMC-Net) to video enhancement tasks. MEMC-Net directly sends the optical flow, interpolation kernels, and contextual features to the frame fusion stage, along with the warped frames, to generate enhanced frames, which produces a fairly large network.

The above analysis shows that compared with single-frame methods, the superior performance of multiframe methods largely benefits from the use of optical flow. Nonetheless, optical flow requires the pixelwise motion estimation and compensation; therefore, good performance has been achieved at the expense of high computational complexity. Besides, more gains may be available if more neighboring frames are involved in the enhancement. But more computational resources are accordingly required. A balanced frame selection strategy is expected, which has not been extensively discussed in the existing work.

To tackle all aforementioned issues, we develop a new multiframe approach for compressed video enhancement, aiming to achieve a balanced tradeoff between enhancement performance and computational complexity. We propose the biprediction-based multiframe video enhancement (PMVE) scheme, in which each current frame is enhanced by the virtual frames (VFs) synthesized from selected frame pairs. Relative to the optical-flow-based method, PMVE completely excludes the use of motion estimation and compensation operations, aiming to achieve superior enhancement performance while maintaining acceptable computational complexity.

The main contributions of this article are summarized as follows.

1) We propose a biprediction-based approach, called PMVE, to enhance the quality of the compressed video. PMVE first identifies the LFs, then creates the VFs through biprediction from selective neighboring frame pairs of the LFs, and finally enhances the LFs using these VFs. Such a new design avoids the explicit use of motion estimation and compensation, but still effectively takes advantage of the cross-frame information. Hence, our PMVE can maintain reasonable complexity while achieving superior performance.
2) We develop a learning-based two-step framework, namely, synthesization and fusion, to effectively implement PMVE. The prediction network (Pred-net), together with a balanced prediction strategy, is proposed for VF synthesization. Afterward, the slowly fused frame-fusion network (FF-net) is utilized to explore the spatiotemporal features jointly.
3) We thoroughly evaluate the performance of our approach. We compare our PMVE against various single-frame and multiframe-based methods. Besides, we implement a high-accuracy optical-flow-based approach (OPT) for comparison. Comprehensive results prove the superiority of our approach in both objective quality and visual quality at a reasonable complexity level.

The remainder of this article is organized as follows. Section II introduces the related work. The motivation behind this article is presented in Section III. Our proposed PMVE approach is depicted in Section IV. The training and test procedure of our approach are described in Section V. Then, Section VI shows our experimental results. Finally, Section VII concludes this article.

II. Related Work

Existing approaches for compressed video enhancement mainly focus on exploiting the pixel correlations within a single frame. To take advantage of the information provided by neighboring frames, multiframe-based enhancement is developed, which is similar to multiframe SR (MFSR) to a certain extent. For example, they both take the degraded frames as input and generate an upgraded counterpart except that video enhancement produces a high-quality version while SR produces an HR version. The challenges they face nonetheless are different. More specifically, compressed video enhancement aims to remove the artifacts introduced by video compression techniques and restore a frame as close to its original version as possible, whereas SR is an ill-posed problem that creates new pixels in the resulting HR version. In this section, after the introduction of single-frame methods, we will further review those approaches of multiframe-based SR and multiframe enhancement, respectively.

A. Single-Frame Quality Enhancement

Encoder Side: Some CNN networks are applied at the encoder side to replace the anchor in-loop filtering algorithm

[16]–[18], or to add a high-dimensional filter to the anchor in-loop filtering [19]–[22], or to switch between the CNN-based and the traditional filters [23]–[29], to further improve the quality of reconstructed frames. Conducted at the encoder side, these methods modify the in-loop filtering algorithms and generate bitstreams that are not standard aligned anymore. This largely prevents such approaches from wide deployment in the real applications, as existing decoders adopted in major video players are, in general, only capable of handling standard bitstreams.

Decoder Side: In contrast, some schemes are proposed to enhance the compressed videos at the decoder side by postprocessing [30]–[33]. They are all single-frame methods, where CNNs are designed to explore only the spatial correlation among pixels within a single frame. The temporal correlation across frames is not utilized, which limits the performance of quality enhancement.

B. Multiframe Super Resolution

The essential idea behind MFSR is to take advantage of the similarities across multiple LR video frames to construct a single HR frame. In the following, we will review different solutions to the MFSR problem.

Traditional Methods: In the early time, Baker and Kanade [41] employed optical flow to model the temporal dependency across frames, proving the effectiveness of optical flow in solving the MFSR problem. Afterward, numerous methods are developed to improve the optical-flow algorithm [42]. Although these algorithms can accurately model the motion across frames, they introduce high computational complexity. Some methods, such as [43] and [44], try to avoid motion estimation by employing nonlocal mean and 3-D steering kernel regression. As a result, the computational complexity is reduced at the expense of frame quality degradation.

CNN-Based MFSR: Lately, many CNN-based MFSR solutions have been proposed. Greaves and Winter [45] skipped the motion estimation step and directly concatenated several frames together for frame fusion. Furthermore, Huang *et al.* [46] employed a bidirectional recurrent convolutional network for SR. Without motion compensation, it is challenging for a single network to sufficiently learn the dependencies across multiple frames. For example, Greaves and Winter [45] found that without dealing with motion, the use of more than two adjacent frames will yield worse results, compared to the use of one adjacent frame. To this end, Kappeler *et al.* [47] adopted a two-step approach where an optical-flow estimation is first applied for frame compensation and then a three-layer video SR network (VSRNet) is designed for frame fusion. Recently, Yi *et al.* [48] proposed to embed the convolutional long short-term memory (ConvLSTM) into ultradense residual blocks (ResBlks) to extract and retain spatiotemporal correlation. Then, a multitemporal information fusion strategy is adopted to merge the temporal feature maps extracted from consecutive frames for SR. Wang *et al.* [49] proposed the multimemory CNN (MMCNN), cascading an optical-flow network that is sped up by a motion transformer operator and an image-reconstruction network with multimemory blocks. These studies take advantage of the spatiotemporal information and achieve significant improvement. It can be seen that motion estimation and compensation are essentially beneficial for promoting the quality of MFSR.

C. Multiframe Video Quality Enhancement

General Purpose Video Enhancement: Lately, video quality enhancement has received significant attention. For example, Wang *et al.* [5] proposed a video restoration framework with enhanced deformable networks (EDVRs), where a pyramid, cascading, and deformable alignment module is devised for frame alignment and a temporal and spatial attention fusion module is employed for subsequent restoration. Bao *et al.* [40] further extended their MEMC-Net framework for video quality enhancement. The neighboring frames are warped through a warping layer, where both optical flow and interpolation kernels are integrated. Afterward, the warped frames, together with the optical-flow results, interpolation kernels, and contextual features, are all fed into the frame enhancement network. Regardless of the quality fluctuation across video frames, EDVR and MEMC-Net directly utilize the prior N and the following N frames to enhance the frame in between. As a result, they perform well, in general, enhancement tasks, such as video SR, denoising, and deblocking, but still leave large room for further improvement specifically on compressed video enhancement.

Compressed Video enhancement: Compressed video enhancement aims to reduce the artifacts introduced by video compression. A natural way is first employing optical flow for motion estimation and compensation and then conducting enhancement. Lu *et al.* [35] followed this way to design a framework, where only the prior frame is involved for the current frame enhancement. Considering that there is frame quality fluctuation in compressed videos, Yang *et al.* [38] and Guan *et al.* [39] designed the MFQE method, which utilizes the peak quality frames (PQFs) in a video to enhance the non-PQFs. MFQE adopts two steps within its framework, that is, motion compensation subnet (MC-subnet) and quality enhancement subnet (QE-subnet). After detecting the PQF, the non-PQF and its two nearest PQFs are fed into the MFQE network, where the motion between the non-PQF and its nearest PQFs is first compensated through the MC-subnet and subsequently the compensated frames and the frame to be enhanced are fused by the QE-subnet. MFQE has demonstrated a higher PSNR performance over most single-frame solutions, but there is still potential to further improve the enhancement performance beyond what MFQE has achieved. For example, two PQFs are involved in enhancing each non-PQF while the non-PQFs near the current frame have been excluded, hence, the cross-frame correlation information may not have been sufficiently utilized. Another potential for improvement lies in the MC-subnet, which employs a lightweighted STMC to estimate the optical flow. Such kind of MC-subnet may not be particularly effective when dealing with those frames with diverse motion and abrupt brightness

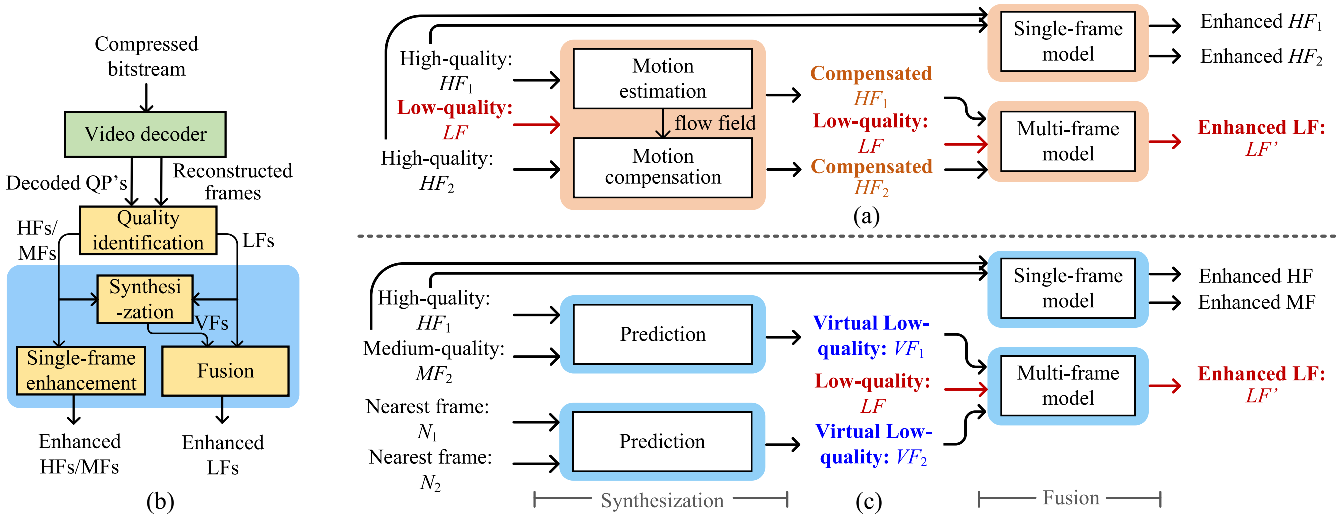


Fig. 1. Proposed learning-based framework. (a) Conventional optical flow-based method. (b) Block diagram of the proposed PMVE approach. (c) Details of the proposed PMVE approach. Here, high-quality frame indicates the HF, medium-quality frame is the MF, low-quality frame is the LF, and VF implies the virtual frame.

change, which might limit the enhancement performance of MFQE.

Different from all the above methods which explicitly utilize optical flow, this article proposes a biprediction-based approach—PMVE, which can generate frames of higher quality without the explicit use of optical flow-based motion estimation and compensation.

III. Motivation Behind This Article

Inspired by the related work, we propose the PMVE scheme for compressed video enhancement. Our motivation behind this article is described as follows.

1) In this article, we target to leverage the potential inherent in the temporal domain. A two-step framework, which has shown its superiority in previous work, is adopted. As described in Fig. 1, the first step is preprocessing and the second is frame fusion. For accurate modeling, both stages within the framework are developed through learning technology.
2) In preprocessing, a method is expected to efficiently extract and collect the temporal information. The previous work generally resorts to optical flow, which can obtain pixelwise motion information. Then, neighboring frames are compensated on the basis of such motion information, as illustrated in Fig. 1(a). Obviously, in such scenarios, the performance largely depends on the quality of optical flow. For accurate modeling, a complicated estimation operation is usually conducted, which requires high computational resources. On the other hand, more neighboring frames are usually involved in video enhancement to achieve better gains. But more computational resources are accordingly required. Essentially, the neighboring frames are similar and abundant redundancy exists in the temporal information they provide, especially for compressed videos where frames have high reference dependencies. Therefore, high performance can be achieved and the complexity can be maintained at a reasonable level if we can extract and utilize the information across multiple frames in an efficient manner. Following the above analysis, we may formulate the information extraction and collection as a single learning operation. Inspired by the techniques behind frame interpolation and extrapolation [40], [50], [51], we propose to extract the temporal information in a biprediction manner, that is, we try to predict the current frame through learning from its prior and following neighboring frames, even without the need of utilizing the current frame. The predicted frame is essentially an inference of the current frame, so called the VF, containing abundant relevant temporal information helpful for the enhancement of the current frame. Relative to the optical flow-based method, the biprediction scheme can involve more neighboring frames for enhancement without increasing the computational complexity.
3) In frame fusion, it is critical to develop a CNN structure taking full advantage of the obtained temporal information. But on the other hand, due to the restriction of available memory, both the depth of CNN and the number of network parameters have to be limited. A large amount of parameters or an unreasonable network structure will also have a large probability leading to the overfitting problem and instead deteriorating the performance. Hence, it is needed to design a CNN approach that can well balance the CNN depth and the total number of network parameters.

IV. Proposed PMVE Approach

A. Framework of PMVE

Fig. 1(b) shows the block diagram of our proposed PMVE approach. After decoding the compressed bitstream, we obtain the reconstructed frames and their base quantization parameter (QP) values. The quality identification module detects the quality of these frames according to their QPs. As in

Algorithm 1: Pseudocode of Our PMVE Approach

```
Input:
  • $F_{i+1}, \cdots, F_{i+L}$: $L$ successive frames in a video
  • $QP_{i+1}, \cdots, QP_{i+L}$: the base QP values of $L$ frames
Output:
  • $EF_{i+1}, \cdots, EF_{i+L}$: the enhanced $L$ frames
Intermediate Variables:
  • $idx\_HF$: frame index of the identified HF
  • $idx\_MF$: frame index of the identified MF
  • $VF_i, (i = 1, 2)$: the predicted virtual frames
Initialization:
  • $idx\_start = 0$
while not at the end of the input video do
    for $(idx{=}idx\_start{+}1;\ idx \leq idx\_start{+}L;\ idx{+}{+})$ do
        $idx\_HF = find\_highest(QP_{idx})$;
        $idx\_MF = find\_second\_highest(QP_{idx})$;
    $EF_{idx\_HF} = SVE\_net(F_{idx\_HF})$;
    $EF_{idx\_MF} = SVE\_net(F_{idx\_MF})$;
    /*Recursively enhance the LFs between HF and MF*/
    $process(idx\_start, idx\_MF)$;
    /*Recursively enhance the LFs between MF and HF*/
    $process(idx\_MF, idx\_HF)$;
    $idx\_start = idx\_HF$;
Function $process(start, end)$
    If $|start - end| <= 1$ then
        return;
    $mid = (start + end)/2$;
    $VF_1 = Pred\_net(F_{start}, F_{end})$;
    $VF_2 = Pred\_net(F_{mid-1}, F_{mid+1})$;
    $EF_{mid} = FF\_net(F_{mid}, VF_1, VF_2)$;
    $process(start, mid)$;
    $process(mid, end)$;
EndFunction
```

Fig. 1(c), only the LFs will proceed through the synthesization and fusion steps, both of which are learning based. All other frames, that is, the HFs and the medium-quality frames (MFs), in contrast, will be enhanced directly by the single-frame video enhancement (SVE). Our SVE consists of one convolutional layer, ten cascading ResBlks, and one output layer (Table I). The pseudocode describing the flow of our approach is provided in Algorithm 1.

B. Frame Quality Identification

In practical video applications, very noticeable quality fluctuation exists across compressed video frames. There are generally several LFs between two HFs. Hence, the HFs can be employed to enhance the LFs since the HFs carry more abundant details, which are helpful for the restoration of LFs. Then, it becomes a problem that how to identify the quality levels of frames in a video.

In this article, we identify the quality level of a given frame from its base QP value. For each compressed frame, a base QP is immediately available after decoding without extra cost. The QP value of each coding block within a frame is allowed to slightly change from the base QP value but not differ significantly. The base QP value hence potentially dominates the quality of a frame. For a group of successive decoded frames, we follow Algorithm 1 to identify the LFs and the HFs/MFs. Starting from the identified LFs, we develop a PMVE approach for multiframe enhancement, as described in the following.

TABLE I
STRUCTURE OF TYPICAL VARIATIONS FOR FF-NET AND OUR SVE NETWORK

Direct fusion: 421k params			Early fusion: 485k params		
Layers	Filter size	Filter number	Layers	Filter size	Filter number
conv 1/2/3	3×3	128	conv 1/2/3	3×3	64
-	-	-	conv 4	1×1	64
ResBlk × 9	1×1 3×3 1×1	64 64 128	ResBlk ×7	1×1 3×3 1×1	64 64 64
conv end	5×5	1	conv end	5×5	1
Slow fusion: 435k params			**SVE: 452k params**		
Layers	Filter size	Filter number	Layers	Filter size	Filter number
conv 1/2/3	3×3	64	conv 1	3×3	64
conv 4/5	1×1	64	-	-	-
conv 6/7	3×3	64	-	-	-
conv 8/9	3×3	64	-	-	-
ResBlk × 5	1×1 3×3 1×1	64 64 128	ResBlk ×10	1×1 3×3 1×1	64 64 64
conv end	5×5	1	conv end	5×5	1

C. Preprocessing

1) Conventional Optical Flow-Based Method: Because there usually exists motion across HFs and LFs in a video, it has been widely adopted to employ a model to explicitly compensate interframe motion. The optical flow could be an ideal candidate. As described in Fig. 1(a), the optical flow-based method, referred to as OPT, first feeds a pair of HFs neighboring to the current LF into the motion estimation network, where two flow fields relative to the LF are generated. Each flow field consists of the pixelwise motion vectors ($\mathbf{MV}_x$ and $\mathbf{MV}_y$) in a frame. Pixels in the two frames are then displaced according to the corresponding flow field. Consequently, we obtain two compensated frames. Suppose that the kth HF is denoted as HF_k and its compensated version is denoted as HF'_k, we can obtain HF'_k through

$$HF'_k = \text{Bilinear}\big(HF_k\big(x + \mathbf{MV}_x, y + \mathbf{MV}_y\big)\big) \tag{1}$$

where $Bilinear(\cdot)$ is the bilinear interpolation function that is employed to deal with the scenario when $(\mathbf{MV}_x, \mathbf{MV}_y)$ are fractional values.

Afterward, the two compensated frames HF'_1 and HF'_2 are sent into the multiframe CNN model, together with the LF to be enhanced.

2) Biprediction Synthesization: As can be seen that the motion estimation and compensation operations in OPT are inefficient. Besides, to enhance an identified LF, only a pair of HFs are utilized, missing the information from neighboring LFs. To address such issues, we propose a biprediction synthesization process. By learning the inherent motion characteristics across frames, the "prediction" module can infer the positions where the objects will occur in the current LF and then produce a VF, namely, VF in Fig. 1. As such, it is essential to design a prediction method for VF inference.

Architecture of Pred-net: In this article, we develop a network called Pred-net to predict the VFs. It adopts a widely used encoder-decoder structure [52]. As shown in Fig. 2, convolutions and downsampling are performed at the encoder and upsampling and convolutions are performed at the decoder.

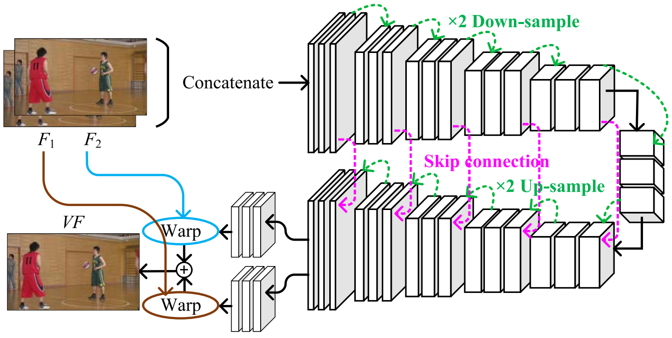


Fig. 2. Proposed Pred-net which adopts the encoder-decoder architecture.

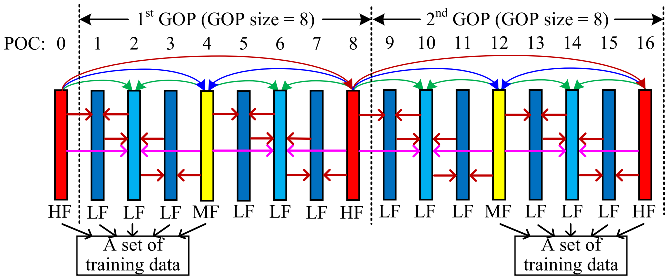

1st GOP (GOP size = 8)
2nd GOP (GOP size = 8)
POC: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
HF LF LF LF MF LF LF LF HF LF LF LF MF LF LF LF HF
A set of training data
A set of training data

Fig. 3. Proposed prediction strategy which is applied to the hierarchical prediction structure of HM. One or two VFs are predicted for each LF to be enhanced. POC is the picture order count and GOP denotes the group of pictures. For example, for LF_1, HF_0 and LF_2 are used to produce a VF, namely, VF_1; for LF_2, HF_0 and MF_4 are employed to generate one VF, namely, VF_21, and LF_1 and LF_3 are used to produce the other VF, namely, VF_22; and in term of LF_3, LF_2 and MF_4 are used to synthesize one VF, namely, VF_3.

Two pixel-adaptive kernels, namely, K_{p1} and K_{p2}, are estimated for the two input frames F_1 and F_2 and the output VF is obtained by summing up their warped results through

$$VF(x,y) = K_{p1}(x,y) * P_1(x,y) + K_{p2}(x,y) * P_2(x,y) \quad (2)$$

where $P_1(x,y)$ and $P_2(x,y)$ denote the patches centered at position (x,y) in frames F_1 and F_2, respectively.

Suppose the size of kernels is S, ideally S^2 parameters are needed for each 2-D kernel. To reduce the memory requirement, we follow the idea in [51] to separate the 2-D convolutional kernels into a pair of 1-D kernels, namely, one vertical kernel and one horizontal kernel. As such, only $2S$ parameters are required for one kernel, largely reduced from the original requirement of S^2. In this article, out of extensive experiments, we empirically set the kernel size as 51. Although some previous work such as Bao *et al.* [40] employed a smaller kernel size than our Pred-net, it additionally integrates optical flow to help handle large motions. On the contrary, our method enlarges the kernel size and produces VFs only through the Pred-net and hence, a relatively larger number is selected.

Balanced Prediction Strategy: As mentioned above, there are generally several LFs between two HFs. The prior work generally employs two HFs to help enhance the LFs in between. Taking the random access (RA) coding scenario of H.265/HEVC reference software HM for example, as described in Fig. 3, the red frames are HFs and the blue ones are LFs to be enhanced. In a group of pictures (GOP), where the picture order count (POC) is numbered from 1 to 8, it is natural to employ HFs of POC = 0 and POC = 8 to enhance LFs labeled from POC = 1 to POC = 7. However, we find that the frame distances between an LF and its two neighboring HFs are usually asymmetrical, indicating that the LF can be of lower temporal correlation with one of the HFs. Meanwhile, it is certain that the LF is in close relationship with its nearest neighbors, although the neighbors are usually of low quality. To cover both frame correlation and frame quality, here we propose a balanced strategy in the biprediction process.

1) We introduce the MF, called MF, to our synthesization for a tradeoff between frame distance and frame quality. Instead of only using HFs to enhance the LFs, MFs are utilized in our synthesization since they are closer to LFs. For example, in Fig. 3, LF_2 is far away from HF_8 but close to MF_4. We hence employ {HF_0, MF_4} instead of {HF_0, HF_8} to enhance the remaining LFs in between.
2) We utilize two pairs of neighboring frames for enhancement. Since the neighboring LFs and the current LF are highly correlated, up to two frame pairs neighboring the current LF are leveraged to predict two VFs. As such, the correlation across frames is explored from two aspects: a) a pair of HF and MF is employed to generate one VF, exploring the benefit from high-quality frames and b) a pair of nearest frames to the current LF are used to synthesize the other VF so as to take full advantage of the information provided from the nearest frames. We take the hierarchical prediction structure in HM as an example to show the proposed selection strategy of frame pairs. As illustrated in Fig. 3, a key frame (red in Fig. 3) together with all the frames between its prior key frame and itself build a GOP. Frames within a GOP are hierarchically predicted. Generally, the frames used as references for the coding of other frames possess higher quality (of smaller QP values) than the other frames, as reference frames of higher quality would usually lead to an overall more ideal rate-quality performance. In Fig. 3, for LF_2, in addition to the VF VF_21 generated from frame pair {HF_0, MF_4}, a pair of LFs {LF_1, LF_3} is employed to produce the other VF VF_22. Both VF_21 and VF_22 are employed to enhance LF_2.
3) But for those LFs whose nearest neighbors already contain HFs or MFs, only the nearest frame pair will be employed for synthesization. For example, in Fig. 3, for LF_3, only a pair of {LF_2, MF_4} are employed. The other frame pair, that is, {LF_1, LF_5} is not only of low quality but also far away from LF_3, which will provide no further benefit for the enhancement of LF_3.

D. Frame Fusion

Given the VFs, the current LF can be enhanced by the frame fusion network FF-net. Generally, there are three ways to deploy the CNN structure of FF-net, including direct fusion, early fusion, and slow fusion. Direct fusion directly collapses all temporal frames together and sends them into a

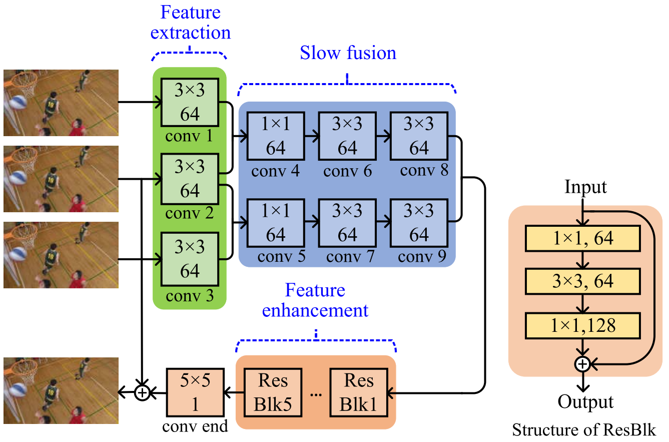


Fig. 4. Structure of our proposed FF-net (left) which adopts a slow-fusion manner, and the detail within one ResBlk (right). There is an ReLU operation after each convolution except the last layer, which is ignored here for simplicity.

TABLE II
PSNR (dB) COMPARISON OF DIFFERENT FF-NET STRUCTURES

Class	H.265/ HEVC	Direct fusion	Early fusion	Slow fusion
A	32.74	33.28	33.28	33.31
B	33.26	33.54	33.54	33.55
C	30.18	30.50	30.50	30.53
D	29.92	30.33	30.32	30.36
E	36.59	37.17	37.16	37.19
Average	32.33	32.73	32.72	32.75

single-frame network. In early fusion, each frame is first processed by a convolutional layer for shallow feature extraction and then the generated feature maps are concatenated to go through a single-frame CNN. Slow fusion merges the temporal information progressively in a hierarchical manner.

On the basis of the residual network [53], we develop three CNN structures, following the above fashions. For a fair comparison, we keep a similar amount of parameters in the three CNNs, as described in Table I. The three CNN models are trained with our training set described in Section V-B and the average performance in 18 common test sequences of H.265/HEVC is provided. Notice that only the first frame of each sequence is tested. We can see from Table II that the slow fusion provides efficiency advantages relative to the early fusion and direction fusion. Our result is in line with the conclusion presented in MFSR, where Kappeler *et al.* [47] and Caballero *et al.* [34] have verified that the slow fusion performs the best. In our experiment, no significant gain is observed from the early fusion compared to direct fusion.

We hence design our FF-net in the slow-fusion fashion, as illustrated in Fig. 4. The first layer is for shallow feature extraction, the next six layers are for slow fusion, and the remaining layers are for feature enhancement. In the stage of shallow feature extraction, two VFs and the current frame are fed into separate convolutional layers. Subsequently, features from two convolutions are concatenated as two parallel branches and go through several convolutions. Then, all channels are gathered to cross five cascading ResBlks for further nonlinear mapping. Finally, the output frame is generated through a 5 × 5 convolution.

V. TRAINING AND TEST PROCEDURES

A. Loss Function

As illustrated in Fig. 1, we denote two VFs output from the Pred-net as $VF_{k,1}$ and $VF_{k,2}$, the current LF as LF_k, and its original frame as LF_k^{GT}. The enhanced LF is represented as LF'_k. Then, the overall loss function of our approach can be described as

$$\mathcal{L}(\theta_{\text{pred}}, \theta_{ff}) = \overbrace{\sum_{i=1}^{2} \alpha_i \cdot \left\| \left(VF_{k,i} - LF_k^{GT} \right) \right\|_1}^{\mathcal{L}_{\text{pred}}} + \overbrace{\beta \cdot \left\| LF'_k - LF_k^{GT} \right\|_2}^{\mathcal{L}_{ff}}. \tag{3}$$

As can be seen that our overall loss function is a linear combination of $\mathcal{L}_{\text{pred}}$ and $\mathcal{L}_{ff}$, which are the losses of Pred-net and FF-net, respectively. During the training, we first set $\beta \to 0$ and $\alpha \to 0.5$ to train the Pred-net. Notice that we expect the Pred-net to concentrate on the task of frame synthesis. If the reconstructed LF is used in $\mathcal{L}_{\text{pred}}$ for training, the Pred-net would learn to handle both encoding artifacts and frame prediction simultaneously, resulting in a training task that is too challenging to obtain an effective network. Therefore, we employ LF_k^{GT} to train the Pred-net, ensuring the prediction procedure can be accurately modeled. Besides, to overcome the oversmoothing problem introduced by the ℓ_2 norm, we adopt ℓ_1 norm in $\mathcal{L}_{\text{pred}}$. When the Pred-net converges, we set the weights $\alpha \to 0$ and $\beta \to 1$ to train the FF-net which uses the ℓ_2 norm optimization.

B. Training Dataset

In our experiments, we use 118 uncompressed video sequences for network training. Each sequence is encoded using the H.265/HEVC reference software HM16.9, at QP values {22, 27, 32, 37}, to obtain the reconstructed videos.

To build the training database, the first 200 frames of each of the 118 sequences are encoded with the default RA configuration *encoder_randomaccess_main.cfg* and low delay P (LDP) configuration *encoder_lowdelay_P_main.cfg* of HM16.9. When selecting frames to train the Pred-net and the FF-net, because adjacent frames are quite similar to each other, an overfitting problem may occur if all frames are contained in the training dataset. To address this issue, we propose a robust training strategy, where the orders of HF and MF are frequently switched for the sake of robustness. As illustrated in Fig. 3, in the first GOP, we select frames from POC = 0 to POC = 4 as a set of training data. In the second GOP, frames from POC = 12 to POC = 16 are selected. By reordering frames from different GOPs in this manner, the trained network will show higher robustness in use.

Furthermore, taking the first GOP in HM as an example, we detail our training strategy. To train the Pred-net, we select four groups of original frames {(frame_0, frame_2),

TABLE III
BD-Rate (%) Comparison With State-of-the-Art Methods (BD-Rate Calculated at QPs 22, 27, 32, and 37)

Class	Sequence	Single-frame methods						Multi-frame methods	
		ARCNN [10]	DnCNN [13]	VRCNN [17]	DSCNN [31]	DCAD [30]	SVE	OPT (0, 4)	PMVE
A	PeopleOnStreet	−4.68%	−9.03%	−6.94%	−8.12%	−10.00%	−9.64%	−11.37%	−11.44%
	Traffic	−5.28%	−8.44%	−7.39%	−8.24%	−9.53%	−9.32%	−10.55%	−10.14%
B	BasketballDrive	−1.77%	−6.41%	−3.85%	−5.77%	−7.79%	−6.98%	−8.43%	−7.94%
	BQTerrace	−2.85%	−8.24%	−6.79%	−8.24%	−10.68%	−9.64%	−10.94%	−10.53%
	Cactus	−4.51%	−8.86%	−7.28%	−8.37%	−10.04%	−9.62%	−10.74%	−10.73%
	Kimono	−2.29%	−5.42%	−4.12%	−5.00%	−6.13%	−5.93%	−7.80%	−7.17%
	ParkScene	−3.55%	−5.79%	−4.95%	−5.62%	−6.34%	−6.15%	−8.09%	−7.49%
C	BasketballDrill	−0.93%	−6.75%	−5.01%	−6.21%	−8.44%	−7.91%	−8.69%	−9.04%
	BQMall	−0.61%	−5.49%	−4.40%	−4.92%	−7.06%	−6.48%	−7.53%	−8.21%
	PartyScene	+4.26%	−2.84%	−1.21%	−1.33%	−3.51%	−2.92%	−4.23%	−4.19%
	RaceHorsesC	−2.42%	−4.87%	−4.01%	−4.56%	−5.58%	−5.30%	−6.69%	−6.30%
D	BasketballPass	−0.39%	−4.77%	−3.94%	−4.73%	−6.11%	−5.78%	−7.05%	−7.87%
	BlowingBubbles	+0.05%	−4.53%	−3.23%	−4.05%	−5.39%	−5.06%	−6.44%	−6.71%
	BQSquare	+7.49%	−6.49%	−2.62%	−3.45%	−7.54%	−6.11%	−8.01%	−7.75%
	RaceHorses	−3.78%	−6.89%	−5.82%	−6.58%	−7.81%	−7.50%	−9.08%	−9.02%
E	FourPeople	−5.93%	−10.79%	−9.49%	−10.99%	−12.91%	−12.74%	−13.19%	−13.21%
	Johnny	−4.24%	−9.68%	−8.52%	−9.81%	−11.64%	−11.15%	−12.12%	−11.97%
	KristenAndSara	−5.40%	−9.46%	−8.76%	−9.77%	−11.49%	−11.36%	−12.46%	−12.21%
	Average	−2.05%	−6.93%	−5.46%	−6.43%	−8.22%	−7.76%	−9.08%	−9.00%

frame_1}, {(frame_0, frame_4), frame_2}, {(frame_1, frame_3), frame_2}, and {(frame_2, frame_4), frame_3} to our dataset. In this way, we establish a training set for the Pred-net, which includes 7600 pairs of training frames.

When the Pred-net converges, we send the selected frame pairs to the Pred-net to predict the VFs. Following the above example, {HF_0, MF_4} and {LF_1, LF_3} are fed into the Pred-net to generate VF_21 and VF_22 for LF_2. A set of training data {VF_21, VF_22, LF_2} is then formed for the FF-net training. As such, a database including 5700 sets of training data is established. Notice that there is only one VF for LF_1 and LF_3. In our training, this VF is repeatedly sent to the FF-net, ensuring that there are always two VFs used for training.

The model trained is employed to enhance LFs. Regarding the HFs and MFs, the single-frame method SVE is applied. To train the SVE model, only HFs and MFs are taken and the resulting dataset includes 3600 frames.

C. Training Settings

Our network is implemented on the Tensorflow platform, using one NVIDIA GeForce GTX 1080Ti GPU for training. During the training, minibatch gradient descent is used for optimization. Frames are segmented into 64 × 64 patches as samples and the batch size is set to 64. We adopt the adaptive moment estimation (Adam) algorithm with an initial learning rate set to 10^{-4}. The learning rate is adjusted using the step strategy with $\gamma = 0.5$. Our validation set includes 50 frames, which are completely excluded from the training database.

D. Test Datasets and Settings

Test Datasets: The 18 test sequences mostly selected by the joint collaborative team on video coding (JCT-VC) for video codec testing and quality assessment are employed for testing. Each sequence is encoded by HM16.9 under the default RA and LDP configurations. Then, the reconstructed frames of each sequence are obtained for further enhancement.

Our test is conducted on a computer with Intel CPU i7-8700@3.20 GHz, 32-GB memory and NVIDIA TITAN V. In the test, we evaluate the average performance of the first 49 frames in each test sequence, including 36 LFs and 13 HFs/MFs. Only PSNR and BD-rate [54] of the luminance component is reported. Details of the code, model, and experimental results can be found in our website [55].

Furthermore, to verify the generalizability of our proposed approach, we employ another two test sets in the experiments of Sections VI-E and VI-H. One dataset is Vimeo90K [56], which is a large-scale video dataset designed for the temporal frame interpolation, video denoising, video deblocking, and video SR. There are 7824 sequences in the Vimeo90K test dataset, and each contains seven consecutive frames with fixed resolution 448 × 256. In this article, we randomly select 900 sequences for performance evaluation. The other dataset contains nine sequences which are usually used for video quality measurement [57]. The nine sequences are structured into four groups according to their resolutions: 1) the 1920 × 1080 group includes *factory*, *life*, and *speed-bag*; 2) the 1280 × 720 group includes *parkrun-ter*, *vidyo1*, and *vidyo3*; 3) the 352 × 288 group includes *bridge-far* and *city-cif*; and 4) the 352 × 240 group includes *garden-sif*.

VI. Experimental Results

Table III describes the overall performance of our approach. For a fair comparison, the SVE network is set with almost the same number of parameters as that of the FF-net of PMVE, as described in Table I. It can be seen that our approach significantly improves the objective quality of compressed frames, which is equivalent to reducing averagely 9.00% BD-rate at the encoder side, compared to the anchor HM16.9.

A. Comparison With Single-Frame Methods

Results: Table III presents the enhancement results of PMVE compared with that of state-of-the-art single-frame methods and our SVE. For a fair comparison, all compared

TABLE IV
PSNR (dB) Comparison With State-of-the-Art Methods (Only LFs)

QP	H.265/HEVC	Single-frame methods						Multi-frame methods	
		ARCNN	DnCNN	VRCNN	DSCNN	DCAD	SVE	OPT (0, 4)	PMVE
22	39.72	39.74	39.94	39.87	39.90	39.95	39.93	39.99	39.97
27	37.20	37.27	37.44	37.39	37.42	37.49	37.46	37.55	37.53
32	34.65	34.73	34.95	34.87	34.91	34.98	34.95	35.05	35.06
37	32.13	32.24	32.43	32.36	32.47	32.49	32.48	32.56	32.58

TABLE V
ΔPSNR (dB) of Our Multiframe Method PMVE and Single-Frame Method SVE

QP	H.265/ HEVC	SVE (ΔPSNR)	PMVE (ΔPSNR)
22	40.22	40.41 (+0.19)	40.44 (+0.22)
27	37.54	37.80 (+0.26)	37.85 (+0.31)
32	34.92	35.22 (+0.30)	35.30 (+0.38)
37	32.33	32.66 (+0.33)	32.75 (+0.42)
42	29.58	29.90 (+0.32)	30.00 (+0.42)
47	27.09	27.38 (+0.29)	27.47 (+0.38)

models are trained with our training database. We can see from Table III that our PMVE surpasses all single-frame methods in the table. Specifically, we show the comparison on 36 LFs in Table IV. It can be seen that PMVE achieves the highest PSNR among all enhancement schemes.

Analysis: Relative to the temporal information, we find that the spatial information contributes more in the enhancement, particularly at small QP values. In Table V, we show the results of ΔPSNR between our multiframe and single-frame methods, compared to the anchor H.265/HEVC. Our SVE adopts almost the same structure and the same number of parameters with the FF-net of PMVE. Hence, the difference between SVE and PMVE is that SVE misses the temporal information derived from the preprocessing stage. We can see that the ΔPSNR improvement comes mainly from SVE, for example, when QP $=$ 22, the gain is 0.19 dB for SVE while 0.22 dB for PMVE, denoting that the temporal domain approximately contributes 0.03 dB. The reason is that the textures of reconstructed frames are well preserved at small QP values so that the artifacts can be easily removed through learning from neighboring pixels. But as the QP value increases, more artifacts are introduced by the encoding process and the contribution of the temporal domain increases. For example, at QP value 42, PMVE achieves an extra 0.10-dB PSNR gain over SVE. But as the QP value increases to 47, the PSNR gain is decreased. The reason is that the correlations in both spatial and temporal domains are so seriously destroyed that it is difficult to achieve further improvement.

B. Comparison With the Optical Flow-Based Method

We compare PMVE with the optical flow-based method, called OPT, which is the most straightforward way for multiframe enhancement. In order to explore the best potential in the temporal domain, we implement an accurate optical-flow algorithm in our experiment, regardless of the computational complexity.

Implementation of the OPT Method: We follow the framework in Fig. 1(a) to implement the OPT method. The compensated frames are generated from motion compensation and sent to the FF-net together with the current LF. Regarding the optical-flow algorithm, FlowNet 2.0 [37] is adopted, as it has been proven in prior work that the accuracy of FlowNet 2.0 is fully on par with state-of-the-art optical-flow methods. Details of the OPT method can be found in our previous work [36]. As such, OPT intensively exploited the temporal correlations. Afterward, the FF-net in PMVE is adopted for frame fusion. Note that here the FF-net is trained for the OPT and our PMVE schemes, respectively. In our experiments, different neighboring frame pairs are utilized in OPT for an extensive comparison.

1) The HF and MF neighboring to the current LF are employed. As described in Fig. 3, {HF_0, MF_4} are used to enhance LF_1, LF_2, and LF_3. This method is called OPT (0, 4).
2) The nearest LFs to the current LF are used, for example, in Fig. 3, {LF_1, LF_3} are employed to enhance LF_2. We term this method as OPT (1, 3).
3) Both the nearest LFs and neighboring {HF, MF} are utilized, which is called OPT $(0, 4, 1, 3)$.

Results and Analysis: We conduct the experiments under the RA configuration of H.265/HEVC. Table VII shows the average PSNR of LFs at QP value 37. As can be seen that OPT (0, 4) performs slightly better than OPT (1, 3), whereas still 0.02 dB lower than PMVE. With respect to OPT (0, 4, 1, 3), because there are four compensated frames input to the subsequent frame fusion stage, the frame fusion network is redesigned to possess the comparable number of parameters to the FF-net. Unfortunately, OPT (0, 4, 1, 3) performs the worst among all. Furthermore, we compare the BD-rate reduction of PMVE and OPT (0, 4) in Table III. We can see that OPT (0, 4) achieves averagely 9.08% BD-rate reduction and our PMVE achieves 9.00%. In general, the enhancement performance of PMVE is comparable to that of OPT.

Although the OPT method is effective, it is computationally expensive. Instead, our proposed approach achieves a comparable enhancement performance with much lower time complexity, as will be presented in Section VI-F.

C. Comparison With State-of-the-Art Work MFQE

We compare PMVE with state-of-the-art work MFQE-1.0 [38] and MFQE-2.0 [39]. For frame quality identification, the same algorithm is applied to PMVE and MFQE. To enhance the identified LFs, we run the open-source code and models provided by MFQE-1.0 [58] and MFQE-2.0 [59] without any modification except the file path. Because MFQE provides only the LDP model at QP value 37, our test is conducted at QP = 37 with default HM16.9 LDP configuration.

To exhibit the performance of both methods accurately, we report only the PSNR of LFs. The results in Table VI show that our PMVE attains an average increment of 0.61-dB PSNR over H.265/HEVC. Compared against MFQE-1.0 and MFQE-2.0, PMVE outperforms 0.40 and 0.09 dB, respectively.

From Table VI, we also observe that PMVE successfully enhances some sequences that MFQE-1.0 fails to deal with. For example, in terms of sequences "PartyScene" and "BQSquare," MFQE-1.0 performs worse than the anchor H.265/HEVC but PMVE reaches as high as 0.49- and 0.88-dB PSNR improvement, respectively. The major reasons are summarized as follows.

1) One reason is that the motions in these sequences are so complex that the MC-subnet of MFQE is unable to accurately estimate such motions. In our proposed method, we adopt two strategies to attain higher accuracy in the temporal domain. One is that we develop the Pred-net to predict VFs for the current LF rather than utilize the optical flow to compensate HFs to the current LF. The other is that we introduce up to two neighboring frame pairs to help the Pred-net collect temporal information, whereas MFQE only employs a pair of HFs. To verify our conjecture above, we experiment on MFQE-1.0 by replacing the MC-subnet with our Pred-net. The new MFQE, called "improved MFQE-1.0" in Table VI, is retrained using our training database. Results show that the improved MFQE-1.0 outperforms the original one by 0.26-dB PSNR. Specifically, for the above sequences "PartyScene" and "BQSquare," it surpasses H.265/HEVC by 0.46- and 0.90-dB PSNR, respectively.
2) Another reason is that the frame fusion network of PMVE, that is, the FF-net, is capable of extracting extensive features. In Table VI, MFQE-1.0 is improved by our Pred-net, indicating that the only difference between PMVE and the improved MFQE-1.0 is due to their frame fusion networks. We observe that PMVE still outperforms 0.14 dB, verifying the effectiveness of our FF-net for frame fusion.

In summary, attributed to the effective biprediction strategy and FF-net structure, PMVE shows advanced versatility and superiority under different scenarios and consistently outperforms H.265/HEVC for all sequences.

TABLE VI
PSNR (dB) PERFORMANCE OF PMVE COMPARED WITH MFQE

Class	Sequence	H.265/ HEVC	MFQE-1.0 [38]	MFQE-2.0 [39]	Improved MFQE-1.0	PMVE
A	PeopleOnStreet	30.85	31.71	31.86	31.87	31.99
	Traffic	33.06	33.52	33.70	33.65	33.72
B	BasketballDrive	33.94	34.09	34.29	34.42	34.45
	BQTerrace	30.79	30.82	31.24	31.24	31.38
	Cactus	31.83	32.23	32.40	32.39	32.49
	Kimono	33.94	34.50	34.56	34.75	34.83
	ParkScene	31.58	32.02	32.14	32.15	32.19
C	BasketballDrill	31.57	31.79	32.15	32.06	32.20
	BQMall	30.23	30.39	30.87	30.75	30.94
	PartykScene	27.10	26.95	27.56	27.41	27.59
	RaceHorses	28.63	28.78	28.99	29.01	29.07
D	BasketballPass	31.07	31.56	32.00	31.76	31.95
	BlowingBubbles	28.61	28.97	29.27	29.13	29.23
	BQSquare	28.09	27.60	28.55	28.50	28.97
	RaceHorses	28.33	28.80	29.02	29.00	29.09
E	FourPeople	34.80	35.35	35.55	35.46	35.64
	Johnny	36.44	36.84	37.09	36.99	37.12
	KristenAndSara	35.93	36.45	36.70	36.56	36.75
	Average	31.49	31.80	32.11	32.06	32.20

D. Comparison for Different Prediction Strategies

As described in Section IV-C, our proposed prediction strategy involves up to two frame pairs to explore the temporal information. To demonstrate the advantage of this strategy, we compare it with the SVE method, the HF/MF method, and the LF-only method. The latter two methods only input two frames for frame fusion: one is the VF generated and the other is the LF to be enhanced. To accommodate the two frames, a frame fusion network that is similar to SVE is trained. The details of the three methods are as follows.

1) The SVE method employs the single-frame model to enhance every single LF.
2) The HF/MF method uses a pair of HF and MF to enhance the LFs.
3) The LF-only method utilizes the LFs nearest to the current LF for enhancement.

Results: It is intuitive that the neighboring HFs will provide more information for the current frame enhancement. Both our previous work [36] and MFQE [38], [39] follow this inference and we indeed receive some gains. To verify the conjecture, we conduct a comprehensive comparison to show the gains from different solutions. As presented in Table VII, surprisingly, in our approach, the HF/MF method contributes equally to the LF-only method at QP value 37, that is, they both achieve 0.38-dB PSNR gain over H.265/HEVC, which is opposed to our initial idea that the HF/MF method would perform much better. Furthermore, our proposed prediction strategy combines both methods together and obtains 0.45-dB gain in total.

Analysis: The results denote that at QP value 37, the LF-only and the HF/MF schemes both provide benefit for the enhancement procedure. More important, the temporal information derived from these two schemes is somewhat different. We therefore can obtain further improvement by jointly utilizing them. For example, both the LF-only and HF/MF methods gain 0.14-dB PSNR over SVE for sequence "PeopleOnStreet." If they provide the same information for LF enhancement, the combination of them would still gain 0.14 dB because no extra information is introduced. However, experimental results show that the combination leads to 0.27-dB PSNR over SVE, that is, there is an additional 0.13 dB increment over the two methods. It implies that the LF-only and HF/MF schemes explore the correlation across frames from different aspects and they both have effect on the enhancement. Our proposed strategy jointly leverages the two frame pairs, receiving more advantages from the temporal domain.

TABLE VII
PSNR (dB) PERFORMANCE OF DIFFERENT PREDICTION STRATEGIES

QP	H.265/ HEVC	SVE	OPT (0, 4)	OPT (1, 3)	OPT (0, 4, 1, 3)	HF/MF method	LF-only method	PMVE
37	32.13	32.48	32.56	32.48	32.51	32.51	32.51	32.58
42	29.58	29.90	-	-	-	29.95	30.00	30.00
47	27.09	27.38	-	-	-	27.46	27.39	27.47

TABLE VIII
PSNR (dB) PERFORMANCE OF PMVE COMPARED WITH MEMC-NET

Test dataset	x264	MEMC-Net [40]	PMVE
JCT-VC 18 sequences	33.19	33.78	34.07
Vimeo90K	34.80	35.69	36.02
9 new sequences*	33.15	33.66	34.02

[*] The 9 sequences are *factory*, *life*, *speed-bag*, *parkrun-ter*, *vidyo1*, *vidyo3*, *bridge-far*, *city-cif*, and *garden-sif*.

But as the QP value increases, the situation changes. At QP value 42, we find that PMVE performs the same as the LF-only method. While at QP value 47, the major contribution to PMVE comes from the HF/MF frame pair, implying that at large QP values, one pair of frames is sufficient for the enhancement task.

E. Comparison With State-of-the-Art Work MEMC-Net

We evaluate the enhancement performance of PMVE, compared to that of MEMC-Net [40]. Three test datasets are employed for evaluation: the 18 sequences from JCT-VC, the Vimeo90K testset [56] that is used in the experiments of MEMC-Net, and the 9 new sequences.

Test Conditions: The open-source code and models of MEMC-Net [60] are utilized for test. All test sequences are encoded with the test conditions provided in MEMC-Net: *libx264* of FFmpeg in default AI configuration and QP value 37. Since MEMC-Net involves seven frames to enhance the middle frame, we encode seven frames in each sequence and only report the quality of the middle frame without loss of generality.

Results and Analysis: From Table VIII, we can see that PMVE averagely outperforms MEMC-Net by 0.33 dB in three test sets. Although both approaches adopt the learning-based framework, they are essentially different.

1) The fundamental flow between MEMC-Net and PMVE is different. MEMC-Net utilizes the optical flow and motion kernels to complete the warping from the current frame to neighboring frames. Then, all warped frames are sent to the frame fusion stage in conjunction with the flow, the motion kernels, and the texture features. As such, optical flow is still required and a large network is leveraged for fusion. In contrast, PMVE utilizes only the motion kernels to synthesize the VFs in relation to the current LF and no explicit optical flow is involved. Besides, only the VFs are used for fusion and the frame fusion network is, hence, much smaller than that of MEMC-Net.
2) The application scenarios of the two methods are different. MEMC-Net aims to solve general video enhancement tasks, such as denoising and deblocking. It is used for intracoded videos without considering the artifacts introduced by intercompression. On the contrary, PMVE is designed to enhance the compressed video quality in intercoding. PMVE considers the frame quality fluctuation artifacts caused by intercoding and proposes the prediction strategy to address such artifacts. It also considers the high frame reference dependencies in intercoding and proposes to take advantage of the VFs for temporal information collection.

F. Comparison on Computational Complexity

The computational complexity of PMVE is evaluated, as shown in Table IX. The average runtime is calculated from sequences where 36 LFs are enhanced. Notice that the runtime provided here excludes the file reading and writing time. We find out that our GPU cannot run the sequences of resolution 2560×1600 with the models provided by MFQE and MEMC-Net, because the required GPU memory exceeds the capability of our GPU. Therefore, we divide each 2560×1600 frame into four parts and process them one by one. The total time is obtained by summing the processing time of four parts.

From Table IX, we can see that the OPT method costs the longest time and leads to extremely high complexity. In PMVE, because up to two frame pairs are processed by separate Pred-nets, the total time is obtained by summing the time cost of two Pred-nets. Nonetheless, the runtime of PMVE is still much less than that of OPT and MFQE-1.0. Compared with MFQE-2.0, although MEMC-Net and PMVE have more parameters, their runtime increment is limited. MEMC-Net runs quite fast except that it requires 4–22 s at the first frame due to GPU model initialization. If more frames are tested, the runtime of MEMC-Net will be further reduced. In general, our PMVE shows great potential to achieve higher enhancement performance over these methods with limited computational complexity increment.

G. Visual Quality Comparison

We illustrate the visual quality of frames from PMVE in Fig. 5. We observe that the frames compressed by H.265/HEVC look artificial. The single-frame methods improve the subjective quality but miss some details, as illustrated in Fig. 5(a). Instead, our proposed PMVE successfully recovers the details, and the enhanced frames look clearer.

In addition, we compare the visual quality of PMVE with that of multiframe methods, including MEMC-Net, MFQE-1.0, and MFQE-2.0, as illustrated in Fig. 5(b) and (c). We find that the frames enhanced by MEMC-Net and MFQE-1.0

TABLE IX
TIME COMPLEXITY COMPARISON (S/FRAME)

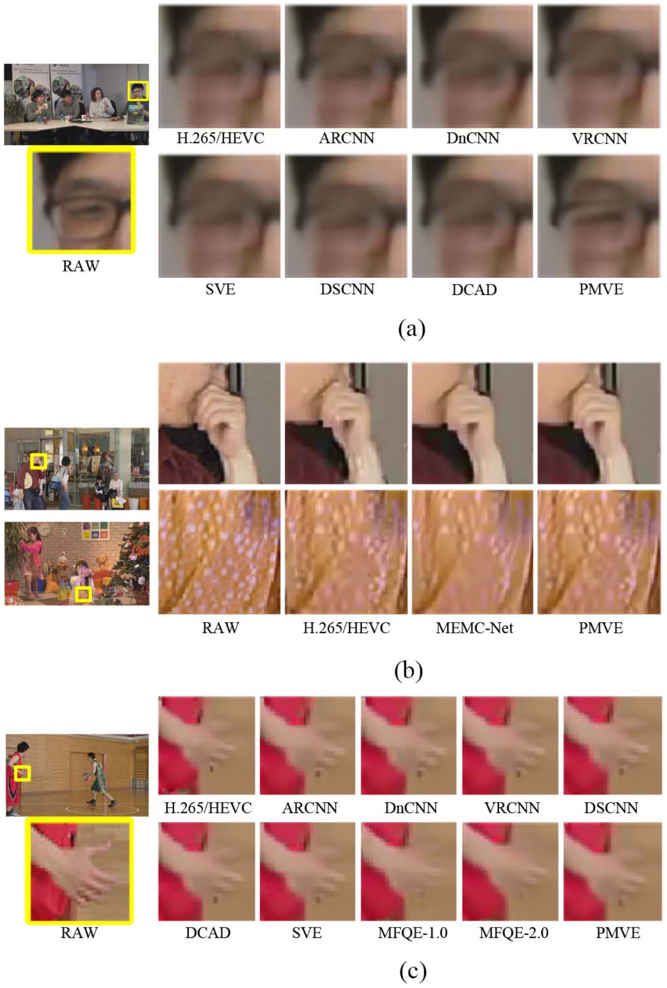
Method	Runtime per frame (seconds)					#Parameters
	2560 × 1600	1920 × 1080	1280 × 720	832 × 480	416 × 240	(million)
OPT	151.414	76.098	38.375	25.645	7.608	160
MFQE-1.0 [38]	2.403	1.158	0.549	0.266	0.097	1.79
MFQE-2.0 [39]	1.298	0.626	0.290	0.128	0.037	0.26
MEMC-Net*[40]	0.420	0.652	0.390	0.190	0.128	72.0
PMVE	1.738	0.873	0.440	0.204	0.069	21.6

* MEMC-Net is run on Pytorch and the other methods are run on Tensorflow.

TABLE X
PSNR (dB) PERFORMANCE OF DIFFERENT METHODS OVER ADDITIONAL TEST DATASETS

Test dataset	H.265/HEVC	Single-frame methods						Multi-frame methods		
		ARCNN [10]	DnCNN [13]	VRCNN [17]	DSCNN [31]	DCAD [30]	SVE	MFQE-1.0 [38]	MFQE-2.0 [39]	PMVE
Vimeo90K	32.86	33.00	33.11	33.07	33.12	33.17	33.17	33.47	33.76	33.81
9 new sequences*	32.00	32.09	32.20	32.17	32.20	32.24	32.21	32.28	32.43	32.53

* The 9 sequences are *factory*, *life*, *speed-bag*, *parkrun-ter*, *vidyo1*, *vidyo3*, *bridge-far*, *city-cif*, and *garden-sif*.



(a)

(b)

(c)

Fig. 5. Visual quality comparison. (a) Compare PMVE with the single-frame methods. (b) Compare PMVE with MEMC-Net. (c) Compare PMVE with MFQE-1.0 and MFQE-2.0.

look blurry, for example, the hand in sequence "BQmall" and "BasketballPass." This phenomenon is slightly improved in MFQE-2.0. Our proposed PMVE successfully recovers certain missed details and the enhanced frames look more visually pleasing.

H. Generalizability of the PMVE Approach

Finally, we verify the generalization capability of our PMVE over additional test datasets. Besides the aforementioned 18 sequences, the enhancement performance of PMVE is also evaluated on Vimeo90K and 9 new sequences, as depicted in Section V-D. The test sequences are all encoded by HM16.9 in LDP configuration at QP value 37. From Table X, we can see that our PMVE surpasses all other methods in PSNR performance, demonstrating the high generalizability of PMVE over different sequences.

VII. CONCLUSION AND FUTURE DIRECTIONS

In this article, we present a new approach, namely, PMVE, to leverage the joint spatiotemporal correlation across frames for the enhancement of compressed videos. For any LF to be enhanced, we propose a biprediction-based scheme, where VFs are first created from the respective neighboring frame pairs through the Pred-net. Specifically, the tradeoff between frame quality and frame distance is considered, and the frame pairs are identified accordingly. Conventional pixelwise motion estimation and compensation process are thus avoided and a large complexity reduction is achieved. Afterward, the VFs are fed into the FF-net for frame fusion, in conjunction with the original LFs to finally reconstruct the enhanced version. The experimental results confirm the effectiveness of our PMVE, as it obtains a consistent superior result in PSNR and visual quality over state-of-the-art work.

Currently, we mainly apply PMVE to the decoder in the postprocessing stage. We are continuing to attempt the use of PMVE at the encoder side. For example, high-quality reference frames may be produced through PMVE and then involved in motion estimation for further coding efficiency improvement. Meanwhile, we will further investigate computational complexity reduction to improve overall performance.

References

[1] T. Hussain, K. Muhammad, J. D. Ser, S. W. Baik, and V. H. C. de Albuquerque, "Intelligent embedded vision for summarization of multiview videos in IIoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2592–2602, Apr. 2020.

[2] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[3] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1646–1654.

[4] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 2472–2481.

[5] X. Wang, K. 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 (CVPRW)*, Long Beach, CA, USA, 2019, pp. 1954–1963.

[6] K. Zeng, J. Yu, R. Wang, C. Li, and D. Tao, "Coupled deep autoencoder for single image super-resolution," *IEEE Trans. Cybern.*, vol. 47, no. 1, pp. 27–37, Jan. 2017.

[7] J. Jiang, Y. Yu, Z. Wang, S. Tang, R. Hu, and J. Ma, "Ensemble super-resolution with a reference dataset," *IEEE Trans. Cybern.*, early access, Mar. 1, 2019, doi: 10.1109/TCYB.2018.2890149.

[8] C. Ren, X. He, Y. Pu, and T. Q. Nguyen, "Learning image profile enhancement and denoising statistics priors for single-image super-resolution," *IEEE Trans. Cybern.*, early access, Aug. 22, 2019, doi: 10.1109/TCYB.2019.2933257.

[9] L. Zhou, Z. Wang, Y. Luo, and Z. Xiong, "Separability and compactness network for image recognition and super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3275–3286, Nov. 2019.

[10] 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)*, Santiago, Chile, 2015, pp. 576–584.

[11] Z. Wang, D. Liu, S. Chang, Q. Ling, Y. Yang, and T. Huang, "D3: Deep dual-domain based fast restoration of jpeg-compressed images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2764–2772.

[12] J. Guo and H. Chao, "Building dual-domain representations for compression artifacts reduction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 628–644.

[13] K. Zhang, W. M. Zuo, Y. J. Chen, D. Y. 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.

[14] K. Li, B. Bare, and B. Yan, "An efficient deep convolutional neural networks model for compressed image deblocking," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Hong Kong, China, 2017, pp. 1320–1325.

[15] Z. Chen, J. Lin, T. Zhou, and F. Wu, "Sequential gating ensemble network for noise robust multiscale face restoration," *IEEE Trans. Cybern.*, early access, Jan. 17, 2019, doi: 10.1109/TCYB.2018.2889791.

[16] W. Park and M. Kim, "CNN-based in-loop filtering for coding efficiency improvement," in *Proc. IEEE 12th Image Video Multidimensional Signal Process. Workshop (IVMSP 2016)*, Bordeaux, France, 2016, pp. 1–5.

[17] 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)*, 2017, pp. 28–39.

[18] J. Kang, S. Kim, and K. Lee, "Multi-modal/multi-scale convolutional neural network based in-loop filter design for next generation video codec," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Beijing, China, 2017, pp. 26–30.

[19] Y. Zhang, T. Shen, X. Ji, Y. Zhang, R. Xiong, and Q. Dai, "Residual highway convolutional neural networks for in-loop filtering in HEVC," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3827–3841, Aug. 2018.

[20] X. Meng, C. Chen, S. Zhu, and B. Zeng, "A new HEVC in-loop filter based on multi-channel long-short-term dependency residual networks," in *Proc. Data Compr. Conf. (DCC)*, Snowbird, UT, USA, 2018, pp. 187–196.

[21] C. Jia *et al.*, "Content-aware convolutional neural network for in-loop filtering in high efficiency video coding," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3343–3356, Jul. 2019.

[22] T. Li, M. Xu, Y. Ren, and X. Tao, "A DenseNet based approach for multi-frame in-loop filter in HEVC," in *Proc. Data Compr. Conf. (DCC)*, Snowbird, UT, USA, 2019, pp. 270–279.

[23] D. Ding, L. Kong, G. Chen, Z. Liu, and Y. Fang, "A switchable deep learning approach for in-loop filtering in video coding," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 15, 2019, doi: 10.1109/TCSVT.2019.2935508.

[24] Y. Hsiao *et al.*, *CE13-1.1: Convolutional Neural Network Loop Filter*, document JVET-N0110, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[25] M. Wang *et al.*, *CE13-Related: In-Loop Filter With Only CNN-Based Filter*, document JVET-N0133, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[26] J. Yao and L. Wang, *CE13-2.1: Convolutional Neural Network Filter (CNNF) for Intra Frame*, document JVET-N0169, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[27] Y. Wang, Z. Chen, Y. Li, L. Zhao, S. Liu, and X. Li, *CE13: Dense Residual Convolutional Neural Network Based in-Loop Filter (Test 2.2 and 2.3)*, document JVET-N0254, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[28] H. Yin, R. Yang, X. Fang, and S. Ma, *CE13-1.2: Adaptive Convolutional Neural Network Loop Filter*, document JVET-N0480, Int. Telecommun. Union, Geneva, Switzerland, 2019.

[29] T. Li, M. Xu, C. Zhu, R. Yang, Z. Wang, and Z. Guan, "A deep learning approach for multi-frame in-loop filter of HEVC," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5663–5678, Nov. 2019.

[30] 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 Compr. Conf. (DCC)*, Snowbird, UT, USA, 2017, pp. 410–419.

[31] 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)*, Hong Kong, China, 2017, pp. 817–822.

[32] X. Song *et al.*, "A practical convolutional neural network as loop filter for intra frame," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, 2018, pp. 1133–1137.

[33] X. He, Q. Hu, X. Han, X. Zhang, C. Zhang, and W. Lin, "Enhancing HEVC compressed videos with a partition-masked convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, 2018, pp. 216–220.

[34] J. Caballero *et al.*, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 4778–4787.

[35] M. Lu, M. Cheng, Y. Xu, S. Pu, Q. Shen, and Z. Ma, "Learned quality enhancement via multi-frame priors for HEVC compliant low-delay applications," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, 2019, pp. 934–938.

[36] J. Tong, X. Wu, D. Ding, Z. Zhu, and Z. Liu, "Learning-based multi-frame video quality enhancement," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, Taiwan, 2019, pp. 929–933.

[37] I. Eddy, M. Nikolaus, S. Tonmoy, K. Margret, D. Alexey, and B. Thomas, "Flownet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 2462–2470.

[38] 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. (CVPR)*, Salt Lake City, UT, USA, 2018, pp. 6664–6673.

[39] 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.*, early access, Oct. 2, 2019, doi: 10.1109/TPAMI.2019.2944806.

[40] W. Bao, W. Lai, X. Zhang, Z. Gao, and M. Yang, "MEMC-Net: Motion estimation and motion compensation driven neural network for video interpolation and enhancement," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 17, 2019, doi: 10.1109/TPAMI.2019.2941941.

[41] S. Baker and T. Kanade, "Super-resolution optical flow," Dept. Robt. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Rep. CMU-RI-TR-99-36, 1999.

[42] C. Liu and D. Sun, "A Bayesian approach to adaptive video super resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, 2011, pp. 209–216.

[43] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the nonlocal-means to super-resolution reconstruction," *IEEE Trans. Image Process.*, vol. 18, no. 1, pp. 36–51, Jan. 2009.

[44] H. Takeda, P. Milanfar, M. Protter, and M. Elad, "Super-resolution without explicit subpixel motion estimation," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 1958–1975, Sep. 2009.

[45] A. Greaves and H. Winter. *Multi-Frame Video Super-Resolution using Convolutional Neural Networks*. Accessed: Apr. 2017. [Online]. Available: http://cs231n.stanford.edu/reports/2016/pdfs/212 Report.pdf.

[46] Y. Huang, W. Wang, and L. Wang, "Bidirectional recurrent convolutional networks for multi-frame super-resolution," in *Proc. 29th Annu. Conf. Adv. Neural Inf. Process. Syst.*, 2015, pp. 235–243.

[47] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE Trans. Comput. Imag.*, vol. 2, no. 2, pp. 109–122, Jun. 2016.

[48] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Jul. 1, 2019, doi: 10.1109/TCSVT.2019.2925844.

[49] Z. Wang *et al.*, "Multi-memory convolutional neural network for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2530–2544, May 2019.

[50] W. Bao, W. Lai, C. Ma, X. Zhang, Z. Gao, and M. H. Yang, "Depth-aware video frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 3703–3712.

[51] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, 2017, pp. 261–270.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Munich, Germany: Springer, 2015, pp. 234–241.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 630–645.

[54] G. Bjøntegaard, "Calculation of average PSNR difference between RD-curves," Int. Telecommun. Union, Geneva, Switzerland, ITU-T VCEG-M33, 2001.

[55] *PMVE Website*. Accessed: Sep. 8, 2019. [Online]. Available: https://github.com/IVC-Projects/PMVE

[56] 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, 2019.

[57] *Video Codec Testing and Quality Measurement*. Accessed: Jan. 25, 2019. [Online]. Available: https://tools.ietf.org/id/draft-ietf-netvc-testing-08.html

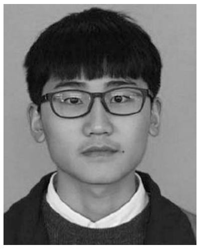
[58] *MFQE-1.0 Website*. Accessed: Jun. 6, 2018. [Online]. Available: https://github.com/ryangBUAA/MFQE

[59] *MFQE-2.0 Website*. Accessed: Aug. 6, 2019. [Online]. Available: https://github.com/RyanXingQL/MFQEv2.0

[60] *MEMC-Net Website*. Accessed: May 6, 2019. [Online]. Available: https://sites.google.com/view/wenbobao/memc-net

Dandan Ding received the B.Eng. (Hons.) and Ph.D. degrees in communication engineering from Zhejiang University, Hangzhou, China, in 2006 and 2011, respectively.

From 2007 to 2008, she was an exchange student with Microelectronic Systems Laboratory (GR-LSM), EPFL, Lausanne, Switzerland. She served first as a Postdoctoral Researcher from 2011 to 2013, then as a Research Associate with Zhejiang University until 2015. Since 2016, she has been a Faculty Member with tenure track with the Department of Information Science and Engineering, Hangzhou Normal University, Hangzhou. She is also with Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking, Hangzhou. Her research interests include artificial intelligence-based image/video processing, video coding algorithm design and optimization, and SoC design.



Wenyu Wang is currently pursuing the bachelor's degree with the Department of Computer Science and Engineering, Hangzhou Normal University, Hangzhou, China.

His research interests include video coding and image processing.

Junchao Tong is currently pursuing the M.S. degree in computer science and engineering with Hangzhou Normal University, Hangzhou, China.

His research interests include video coding/processing and machine learning.

Xinbo Gao (Senior Member, IEEE) received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively.

From 1997 to 1998, he was a Research Fellow with the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Postdoctoral Research Fellow with the Department of Information Engineering, Chinese University of Hong Kong, Hong Kong. Since 2001, he has been with the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor with the Ministry of Education, a Professor of pattern recognition and intelligent systems, and the Dean of the Graduate School, Xidian University. He is also with the Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China. He has published six books and around 300 technical articles in refereed journals and proceedings. His current research interests include image processing, computer vision, multimedia analysis, machine learning, and pattern recognition.

Prof. Gao is on the editorial boards of several journals, including *Signal Processing* (Elsevier) and *Neurocomputing* (Elsevier). He served as the general chair/co-chair, the program committee chair/co-chair, or the PC member for around 30 major international conferences. He is a fellow of the Institute of Engineering and Technology and the Chinese Institute of Electronics.

Zoe Liu received the B.E./M.E. degree from Tsinghua University, Beijing, China, in 1995 and 2000, respectively, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2004.

She was a Software Engineer with Google WebM Team, Mountain View, CA, USA, and has been a key contributor to the royalty free video codec standard AOM/AV1. She is the Co-Founder and the President of Visionular Inc., Mountain View, a startup delivering cutting-edge video solutions to enterprise customers worldwide. Her main research interests include image/video processing and machine learning.

Yong Fang received the B.Eng., M.Eng., and Ph.D. degrees from the School of Communications Engineering, Xidian University, Xi'an, China, in 2000, 2003, and 2005, respectively.

In 2007, he was a Lecturer with the School of Electronics and Information Engineering, Northwestern Polytechnic University, Xi'an. From 2007 to 2008, he was a Research Professor with the Department of Electrical and Computer Engineering, Hanyang University, Seoul, South Korea. From 2009 to 2016, he was a Full Professor with the College of Information Engineering, Northwest A&F University, Xianyang, China. He is currently a Full Professor with the School of Information Engineering, Chang'an University, Xi'an. He has a lot of experience in hardware development, for example, field-programmable gate-array-based (Xilinx Vertex series) video codec design, and digital-signal-processing-based (TI C64 series) video surveillance systems. His research interests include information theory, pattern recognition, compressive sensing, image/video coding, processing, and transmission.

Dr. Fang is an Editor of IEEE COMMUNICATIONS LETTERS.