IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 October 2023, accepted 26 October 2023, date of publication 2 November 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3329949

RESEARCH ARTICLE

Robust Spatial-Temporal Motion Coherent Priors for Multi-View Video Coding Artifact Reduction

GYULEE JEON[1], YEONJIN LEE[1], JUNG-KYUNG LEE[2], YONG-HWAN KIM[3], AND JE-WON KANG[1,2], (Member, IEEE)
[1]Graduate Program in Smart Factory, Ewha Womans University, Seoul 03760, South Korea
[2]Department of Electronic and Electrical Engineering, Ewha Womans University, Seoul 03760, South Korea
[3]Korea Electronics Technology Institute, Seongnam-si 13509, South Korea

Corresponding author: Je-Won Kang (jewonk@ewha.ac.kr)

This work was supported in part by NRF funded by MSIT under Grant NRF-2022R1A2C4002052; in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) funded by the Korean Government (MSIT); in part by the Development of Ultra High Resolution Unstructured Plenoptic Video Storage, Compression, and Streaming Technology for Medium to Large Space under Grant 2020-0-00920; and in part by the Development of Moving Robot-Based Immersive Video Acquisition and Processing System in Metaverse under Grant RS-2022-00167169.

ABSTRACT Multi-view video (MVV) data processed by three-dimensional (3D) video systems often suffer from compression artifacts, which can degrade the rendering quality of 3D spaces. In this paper, we focus on the task of artifact reduction in multi-view video compression using spatial and temporal motion priors. Previous MVV quality enhancement networks using a warping-and-fusion approach employed reference-to-target motion priors to exploit inter-view and temporal correlation among MVV frames. However, these motion priors were sensitive to quantization noise, and the warping accuracy was degraded, when the target frame used low-quality features in the corresponding search. To overcome these limitations, we propose a novel approach that utilizes bilateral spatial and temporal motion priors, leveraging the geometry relations of a structured MVV camera system, to exploit motion coherency. Our method involves a multi-view prior generation module that produces both unidirectional and bilateral warping vectors to exploit rich features in adjacent reference MVV frames and generate robust warping features. These features are further refined to account for unreliable alignments cross MVV frames caused by occlusions. The performance of the proposed method is evaluated in comparison with state-of-the-art MVV quality enhancement networks. Synthetic MVV dataset facilitates to train our network that produces various motion priors. Experimental results demonstrate that the proposed method significantly improves the quality of the reconstructed MVV frames in recent video coding standards such as the multi-view extension of High Efficiency Video Coding and the MPEG immersive video standard.

INDEX TERMS Multi-view video compression, video enhancement, motion vector, VVC, MPEG-immersive video, TMIV.

I. INTRODUCTION

Multi-view video (MVV) has been widely used for emerging three-dimensional (3D) video systems and services such as metaverse and virtual reality (VR), providing a user with a more augmented experience. A VR user wearing a head-mount device can freely navigate a 360-degree virtual space and enjoy a video content in an arbitrary viewpoint with 6 degree-of-freedom (DoF) [3], [4]. For these 3D video systems, the high quality of MVV is required to faithfully realize a 3D space and render acceptable quality of synthesized viewpoints [5]. However, transmitting the large number of views causes significant network bandwidth problems. The quality of video data degrades due to quantization noise after compression, producing visually unpleasant artifacts such as blocking artifacts and blurring effects.

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Montalban.

In the past decades, 3D video coding standards have been established to meet the quality of experience [1], [2], [6] based upon existing 2D video coding standards [7], [8]. When compressing MVVs, inter-view correlation has been exploited in addition to spatial and temporal correlation [9], [10]. Although the 3D video coding standards have attempted to reduce the size of MVV data, they had the same drawbacks as in the conventional lossy coding system. While viewpoints are generated in a 3D virtual space using texture and geometry components, the distortion would cause severe VR sickness during a navigation [11].

Several post-processing techniques have been developed to restore the quality of MVV data. Convolutional neural network (CNN) has been actively used to develop a compression artifact reduction network (CARN). The CARNs were used to remove unwanted artifacts from compressed images [12], [13], [14] and videos [15], [16], [17]. Because MVV data are obtained from cameras that are located on the horizontal and vertical planes, inter-view correlation could be further exploited for artifact reduction [18], [19]. In [20], a multi-view denoising network (MVCNN) was presented to manage 3D focus image stacks and fuse them to compute an enhanced target image. The MVCNN was trained with residual learning and batch normalization on top of a denoising CNN [21]. In [22], the same baseline was resorted to reduce artifacts in stereo images. Conventional CARN methods have attempted to use richer appearance features in spatially and temporally adjacent reference frames [18], [23], [24]. In [23] and [24], channel attention mechanisms have been exploited to choose relevant features in reference frames. In [18], a cross-scale warping module based on a spatial transformer network (STN) has been introduced to use the spatial priors of adjacent light filed images. The reference and the target images were aligned with a warping vector to transfer the high quality of reference features. It demonstrated that transferring inter-view features could improve the quality of a target view. However, their unidirectional optical flow was sensitive to quantization noise, which would readily mislead spatial priors after warping.

This paper focuses on utilizing robust spatial and temporal priors for compression artifact reduction of MVV frames. To this aim, we develop a multi-view prior generation (MPG) module to improve the performance of an MVV CARN. In previous works [25], [26], considering an MVV compression circumstance, a current frame is coded using a high quality of reference frames that are coded with lower quantization parameters (QPs) of lower temporal layers [27]. The prediction structure allowed for the current frame to employ the high quality of spatial and temporal reference features along camera and temporal directions, respectively. However, the previous works had limitations to use such MVV prediction structures [4], [25], and we believe there is more room to exploit the structure. In [25], multi-view image quality enhancement (MVIQE) network has been developed to produce a spatial prior using only the disparity vector.

In this paper, learning to convey relevant spatial and temporal features from the reference to the current frame is carefully investigated. While a warping-and-fusion method based on a unidirectional warping vector is employed in our previous work [25], a bilateral warping vector estimation and fusion method is further introduced to consider the geometry relations of MVV camera systems and generate robust warping features. These features are further refined to remedy unreliable alignments cross MVV frames caused by occlusions. Various warping vectors along the camera and temporal directions are included to improve performance.

Motion prediction has played an important role in video generation tasks, including view synthesis [28] and video frame interpolation and extrapolation [29], [30], [31]. However, it was not fully investigated in an MVV compression circumstance due to the insufficient number of video sequences. The MVV test sequences have been actively used for evaluating coding performance in 3D video standardization [1], [2], but a larger size of MVV dataset are required to train a reference-based compression artifact reduction, using spatial and temporal correlation in 3D video coding. We generate computer-synthesized MVV frames for training to challenge this problem.

Our primary contributions are described as follows:

- We propose an MPG module to generate coherent spatial disparity and temporal motion priors to improve the performance of compression artifact reduction for MVV frames. This is accomplished by using a bilateral vector cross MVV frames that are captured from a structured MVV camera system and compressed with 3D video coding standards. Our network includes a refinement module to remedy unreliable alignment caused by occlusions and temporal inconsistency.
- Experimental results demonstrate that the proposed method outperforms state-of-the-arts studies in MVV compression artifact reduction. Our model is tested with recent 3D video coding standards such as the multi-view extension of High Efficiency Video Coding [1] and the MPEG immersive video standard [2].

II. RELATED WORKS

A. PREVIOUS CARNS FOR SINGLE AND MULTI-VIEW VIDEO

Several single-image-based CARNs have been developed to improve the quality of compressed images. In early studies, Zhang et al. [21] and Ehrlich et al. [13] presented denoising CNNs for image restoration for JPEG-compressed images. CNN-based models were applied to compressed images using video coding standards. Dai et al. [32] and Zhang et al. [33] developed deep residual CNNs using variable filter sizes to remove the quantization noise of High Efficiency Video Coding (HEVC)-compressed frames. Generative adversarial network (GAN) was used to generate an improved quality of an image [34]. They used a single input image to improve. However, the methods cannot be efficiently applicable to

MVVs, because it is difficult to exploit inter-view correlation. Multiple images were used to provide more priors, and their features were digested through various fusion methods for single and multi-view videos. Zhu et al. proposed a CNN-based post-processing module to improve the quality of synthesized view and reduce warping distortion in 3D-HEVC [35]. Input frames were concatenated and fused in a network. Jammal et al. used low-quality and high-quality images from different views of the same scene to produce an enhanced image [36]. In [37], multi-view graph neural network (MV-GNN) was proposed to alleviate quantization noise of a target image.

The performance of a CARN can be further improved by transferring useful features from a source to a target [25]. For this purpose, a warping module has been introduced to search a rich feature in a reference frame and locate it to the corresponding position of the current frame. There were several reference-based image quality enhancement studies using a warping-and-fusion approach [18], [19]. The methods have been originally designed for super-resolution (SR) and extended to artifact reduction (AR) later. When a high-quality image was used for a reference, a target image could exploit the corresponding patch to improve the quality in feature levels. CrossNet [18] has been developed to use a cross-scale optical flow to fuse features from light field images. TTSR (Texture Transformer Network) [19] was also difficult to apply to reconstructing compressed contents due to the loss of texture information. Although CrossNet and TTSR has attempted to utilize reference frames for image quality enhancement, since they considered limited number of reference frames, it does not use the multi-view characteristics sufficiently.

There are several studies to use multiple reference images with a warping module for MVV frames. A recent video super resolution method with recurrent back-projection network (RBPN) [38] attempted to extract residual features from the neighboring frames to develop multi-image super resolution (MISR) method. While the RBPN enhances the quality of the target frame by adding missing details from neighboring temporal frames, it does not aim to reduce compression artifacts and can only consider low quality temporal frames of the corresponding target video. Yang et al. presented a multi-frame quality enhancement (MFQE) method to exploit the correlation among input video frames [26], [39]. They selected several frames in a video sequence by using a convolutional long short-term memory (ConvLSTM) module, because the quality of video frames could be fluctuated with a different QP value, determined by a coding configuration. In this way, it can exploit rich texture video patches from neighboring frames, when enhancing the target frame. However, this method only exploits the frames in the same target video sequence and does not consider other useful information from other views or neighboring frames. Furthermore, they used the same unidirectional warping vector, which would degraded the warping accuracy. Lu et al. proposed a quality enhancement network (QE-Net) for an HEVC low-delay configuration. They used a fusion model of spatial and temporal features extracted from multi-scaled convolution layers [40]. Chen et al. presented a residual network to remove quantization artifacts of multi-view depth videos [41]. However, it was indicated that inaccurate warping vectors could transfer undesirable priors [37].

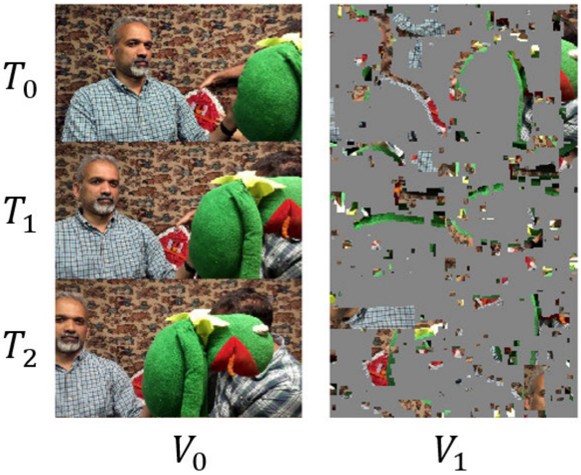


FIGURE 1. MPEG-Immersive video (MIV) compression using atlas patches for base (V_0) and additional views (V_1) in the time steps of T_0, T_1, and T_2. Residual patches generated from a pruning process of additional views contain high-frequency components, which would suffer substantial degradation of visual qualities due to compression artifacts.

Lastly, MVIQE [25] enhances the target image by warping spatial reference images using disparity vectors estimated by PWC-Net [42] which is a widely used flow estimation network. However, because MVIQE only uses unidirectional motion vectors, it is also prone to performance degradation with quantization noise. Furthermore, it does not consider temporal correlation between consecutive frames and therefore is not suitable for video artifact reduction.

B. MVV COMPRESSION

The 3D video coding standards in ISO/IEC Moving Picture Expert Group (MPEG) have been established to reduce the size of MVV data [1], [2] based upon existing 2D video coding standards such as HEVC [7] and Versatile Video Coding (VVC) [8]. In the multi-view HEVC (MV-HEVC) and 3D HEVC standards [1], inter-view redundancy was reduced using a disparity vector (DV), because the positions of the multi-view cameras are structured in an 1-D arc or a 2-D grid coordinate. A DV was used for searching a corresponding block among views and subsequently applied to conventional motion and disparity compensated prediction of a conventional codec pipeline [9], [10].

MPEG-Immersive video (MIV) group [2] has developed a codec-agnostic approach to manage a number of input views for MVV compression. MVVs are pre-processed to create atlas patches to contain common data patches among input views and residual data patches after inter-view prediction. For example, as shown in Fig.1, all the samples from base views are packed into an atlas V_0, whereas samples from additional views are pruned and packed into the other atlas V_1. The pruning process is conducted with an adjacent reference view. The atlas patches are then compressed, using a legacy 2D video codec. In MIV reference software (TMIV), the residual patches generated from a pruning process of additional views contain high-frequency components, which would suffer substantial degradation of visual qualities, as shown in Fig.1. Furthermore, it contains synthesis artifacts.

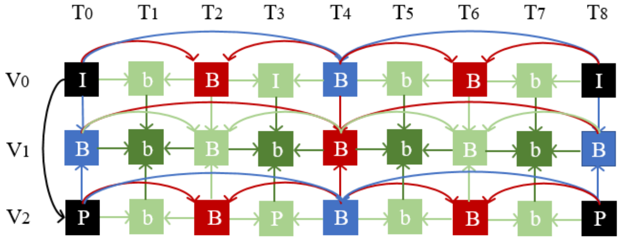


FIGURE 2. MV-HEVC HBP coding structure for MVV coding. V_i and T_i denote indices of viewpoints and time steps, respectively.

In the MVV coding standards, a base-view, which is denoted by V_0 in Fig. 2, is compressed using only the temporal correlation with previously coded frames at different time steps. While a base-view is independently coded with the other views, non-base views are compressed using not only temporal prediction but also inter-view prediction. For example, in Fig. 2, the frames of V_1 are coded using inter-view reference frames of V_0 and V_2 at the same time step and temporal reference of the same viewpoint. The HBP coding structure in Fig. 2 presents five layered groups with different colors. QP values are determined by the layers. I_0 in the base-view is coded using the lowest QP value, which produces the highest peak signal-to-noise ratio (PSNR) value among all the MVV frames. The frames in the next layer (i.e. B frames colored with blue) are coded with reference frames from the lower layer, using an increasing offset to the QP. In Fig. 2, both inter-view prediciton and temporal prediction are applied to a B frame at T_4 and V_1 in the third layer, by referencing the B frames of the second layer.

Because the non-base view frames are coded with higher QP values than in the base view frames, the quality of the non-base view frames drop with the HBP coding structure of MV-HEVC and 3D-HEVC. In TMIV, the atlas patches often contain high-frequency components as shown in Fig. 1, when they are generated from a pruning process. However, many details in the patches would be blurred due to quantization, which degrades the quality of the non-base views.

Multi-view plus depth (MVD) format video data, used for the 3D video coding standards, is represented with a pair of multi-view texture and depth videos. Depth videos are captured from depth camera directly or derived from stereo matching of the corresponding texture video data. Yu et al. [43] used depth video data to locate the high-quality texture to display a rich virtual view using the depth image-based rendering (DIBR) method and improve the low-quality view. However, high-quality and high-resolution depth video format data is not usually available in a decoder side due to a inaccurate depth sensor. An intermediate view can be generated using DIBR techniques in 3D video coding. Because the depth information is usually incomplete, the quality of an intermediate view also drops dramatically. TMIV reference software can drop a depth map and let a decoder generate it to improve overall coding performance [44].

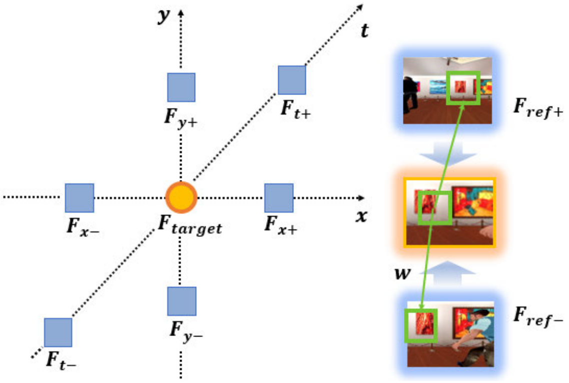


FIGURE 3. MVV frames captured from 2D camera arrays and adjacent spatial and temporal reference frames to enhance the current frame, using unidirectional reference-to-target warping vectors.

III. PROPOSED METHOD

A. MOTIVATION AND PROBLEM FORMULATION

In 3D video compression, a reconstructed frame exhibits quantization artifacts with a high QP value, which significantly degrades visual quality. This study proposes a method to enhance the quality of the current frame F_{tar} using a set of adjacent spatial and temporal reference frames $F_r = \{F_{x+}, F_{x-}, F_{y+}, F_{y-}, F_{t+}, F_{t-}\}$ with high fidelity in MVV compression. MVV frames are captured in 2D camera arrays, as shown in Fig. 3. F_{x+} and F_{x-} represent MVV frames of the left and right views along an x-axis, respectively. Similarly, F_{y+} and F_{y-} represent the MVV frames aligned to an y-axis. F_{t+} and F_{t-} are the forward and backward temporal reference frames. In the 3D video coding standards, because the reference frames are coded with higher quality in an HBP coding structure, they can yield useful priors to enhance the quality of the current frame. Following this idea, our goal is formulated by,

$$F_{enh} = \mathcal{N}(F_{tar}, F_r, w \mid \theta), \tag{1}$$

where $\mathcal{N}$ is a model, F_{enh} is the enhanced output of F_{tar}, w is a learned warping vector, and θ is a set of learnable parameters.

In (1), we adopt a warping-and-fusion approach as in MVIQE [25]. In the method, a unidirectional vector w is trained to warp a reference feature to a current feature, and a warped feature is used as priors to characterize a disparity between a reference frame and a current frame. In our observation, MVV frames tend to display symmetric warping vectors from the current patch to the corresponding

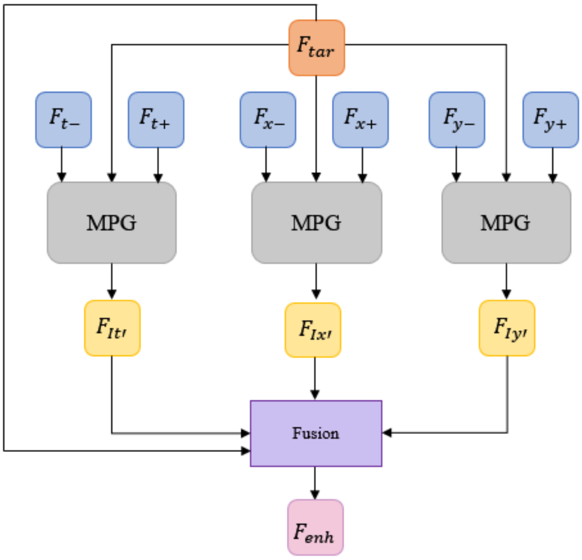


FIGURE 4. F_{tar}, F_{r+}, and F_{r-} are used to derive unidirectional and bilateral warping vectors, i.e., w^u and w^b, and produce an intermediate frame $F_{I'}$. MPG is used to generate the intermediate frames in an x-axis, y-axis, and a temporal direction. The final output is produced from the intermediate frames by fusion.

patches of two reference frames in the both direction, i.e. $w_{r+} \simeq -kw_{r-}$, and k is approximately one, when the frames are captured from a structured camera array. The priors are useful because the bilateral vector is robust to noise for frame generation [45]. Therefore, we use both a unidirectional warping vector as in the MVIQE and a bilateral warping vector as complements and consider them in our network design, given as

$$F_{enh} = \mathcal{N}(F_{tar}, F_{r+}, F_{r-}, w^b_{r\pm}, w^u_{r\pm} \mid \theta), \tag{2}$$

where F_{r+} and F_{r-} are the two reference frames in the positive and negative directions, and w^b and w^u are a bilateral and a unidirectional warping vector, respectively.

Furthermore, we consider a network to learn a warping vector for each axis. Fig. 4 presents an overall view of the network architecture with the MPG modules, individually using F_x, F_y, F_t, and F_{tar}. The three MPGs share the same network architecture but differently trained. The network produces intermediate frames $F_{Ix'}$, $F_{Iy'}$, and $F_{It'}$ to fuse the final output, using (2).

B. MULTI-VIEW PRIOR GENERATION MODULE

In this subsection, we explain an MPG module to produce the priors of the current frame F_{tar}, by applying warping vectors to two reference frames F_{r+} and F_{r-}, as shown in Fig. 5. The proposed network first extracts feature vectors f_{tar}, f_{r+}, and f_{r-} to perform the task in a feature domain. Then, unidirectional warping vectors and bilateral warping vectors are used to produce the priors.

1) UNIDIRECTIONAL MOTION-BASED PRIOR GENERATION

In the MPG, PWC-Net [42] is applied to derive a forward and backward unidirectional warping vectors, denoted by w_{r+} and w_{r-}, as in [25]. The vectors are used to warp the features of the reference frames. In Fig. 5, a reference feature map f_{r+} is warped into f^u_{r+}, given as

$$f^u_{r+}(z) = f_{r+}(z + w^u_{r+}), \tag{3}$$

where w^u_{r+} represents a forward warping vector. Similarly, we compute a backward warping vector w_{r-} to produce f_{r-}, given as

$$f^u_{r-}(z) = f_{r-}(z + w^u_{r-}). \tag{4}$$

The two unidirectional warping vectors have been used to generate the corresponding intermediate frames.

2) BILATERAL MOTION-BASED PRIOR GENERATION

Unidirectional vectors are tend to be sensitive to quantization noise in MVV compression, because a target frame suffers from more severe degradation and a reference-to-target prediction leads to inaccurate warping textures. The network has attempted to tackle quantization noise, but unreliable warping vectors would cause distortion.

Therefore, considering the geometry relations of MVV camera systems [46], we include another prior generated from a bilateral warping vector w^b_r. As shown in Fig. 6, w^b_r presented with a blue solid line is chosen among bilateral vectors around a current frame. Such approach that involves reference-to-reference prediction tends to exhibit robustness to quantization noise due to fine quantization step size of reference frames. The vector is obtained by minimizing the difference between f_{r+} and f_{r-} around the current pixel z, given as

$$w^*_s = \arg\min_{w_s \in S} |f_{r+}(z + \frac{w_s}{2}) - f_{r-}(z - \frac{w_s}{2})|^2_2, \tag{5}$$

where S is a search range of $[-s, s] \times [-s, s]$. Because PWC-Net [42] has been used to compute an optical flow using five-layered feature pyramid, s is adjusted to 6, 4, 4, 2, and 2, in each layer.

Then, we calculate w^b_{r+} and w^b_{r-} as $w^*_s/2$ and $-w^*_s/2$, respectively. The intermediate frames are generated by,

$$f^b_{r+}(z) = f_{r+}(z + w^b_{r+}), \tag{6}$$

and

$$f^b_{r-}(z) = f_{r-}(z + w^b_{r-}), \tag{7}$$

where w^b_{r+} and w^b_{r-} represent a forward and a backward bilateral warping vector. In the next subsection, We explain the derivation method of w^b_{r+} and w^b_{r-}.

The bilateral warping vectors are used to transfer the corresponding texture of the reference frames captured from a structured MVV camera array. However, the accuracy would be degraded with the positions of objects relative to MVV cameras, although the camera system is calibrated appropriately [46]. As a remedy, we modify the bilateral

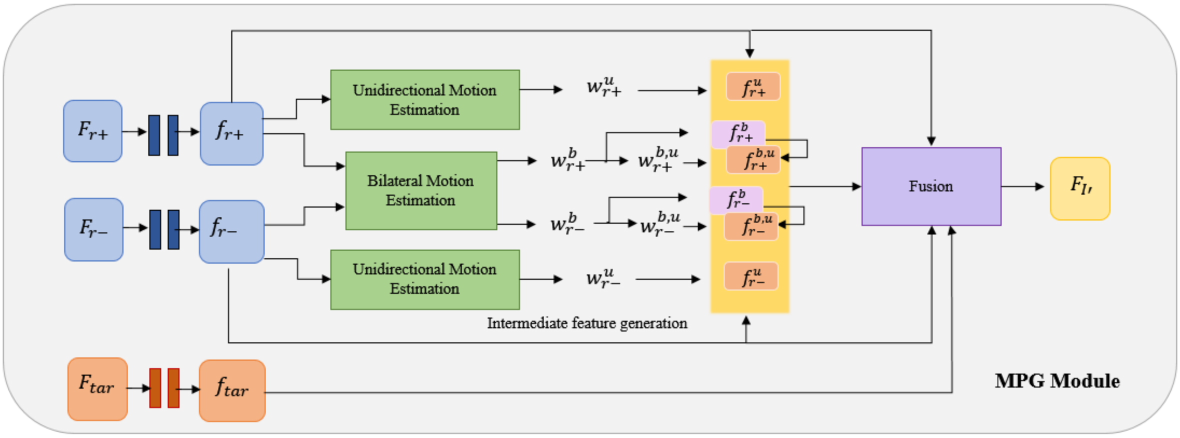


FIGURE 5. Multi-view prior generation module to produce two unidirectional warping vectors and bilateral warping vectors as complements to estimate accurate priors. All warped frames are fused to produce the final output.

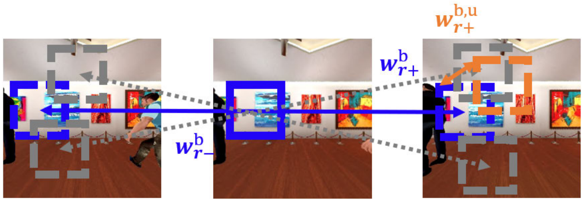


FIGURE 6. Illustration of a bilateral warping vector and its refinement.

vectors, using a two-stage generation method. We also explain the procedure in the following.

3) IMPLEMENTATION OF BILATERAL MOTION ESTIMATION AND REFINEMENT MODULE

We explain the implementations of bilateral motion estimation and refinement module in Fig. 5 to optimize (5), and produce w_{r+}^{b} and w_{r-}^{b}. The overall procedure is presented Fig. 7. The bilateral motion estimation generates an L-level feature pyramid for both reference features f_{r+} and f_{r-} along each axis. As in [42], a warping vector $w^{b}(l-1)$ extracted from an $l-1$ level feature is used to calculate the current warping vector $w^{b}(l)$ in the next level. We adopted a warping layer and a bilateral cost volume layer of [31] in the architecture of bilateral motion estimation. Motivated by [47], the features of reference frames are warped and used to produce the current warping vector through a correlation layer with a bilateral frame and subsequent convolution layers. We use five convolution layers, in which each of output channels is 128, 128, 96, 64, and 32 with the kernel size of 3. Then, the output features and warping vectors are up-sampled for the next level, by using convolution layers with kernel size of 4. This procedure is repeated for all levels of the feature pyramid, and, when $l = 6$, the final warping vector w_{r+}^{b} is produced. While w_{r+}^{b} is the forward bilateral warping vector between the target frame and the forward reference frame, The backward bilateral warping vector w_{r-}^{b} is obtained by reversing w_{r+}^{b}, symmetrically.

In Fig. 7, w_{r+}^{b} and w_{r-}^{b} are further refined to $w_{r+}^{b,u}$ and $w_{r-}^{b,u}$, respectively. For this, w_{r+}^{b} and w_{r-}^{b} are first used to generate $f_{r+}^{b}(z)$ and $f_{r-}^{b}(z)$ as intermediate features using (6) and (7), respectively. Then, f_{r+}^{b} and f_{r-}^{b} are used as reference features to obtain $w_{r+}^{b,u}$ and $w_{r-}^{b,u}$, respectively. That is mathematically defined as,

$$w_{r+}^{b,u*} = \arg\min_{w_{r+}^{b,u}\in S} |f_{r+}(z) - f_{r+}^{b}(z + w_{r+}^{b,u})|_2^2, \tag{8}$$

and

$$w_{r-}^{b,u*} = \arg\min_{w_{r-}^{b,u}\in S} |f_{r-}(z) - f_{r-}^{b}(z + w_{r-}^{b,u})|_2^2, \tag{9}$$

where S is a search range as in (5). The optimization is conducted using a flow refinement module colored with orange in Fig. 7, in which a bilateral frame prior is first generated as in the bilateral motion estimation module and used to compute the difference with the references. The subsequent convolution layers are used to calculate the refined warping vectors. The sizes of the kernels are 3, while the last convolution layer used for up-sampled features uses the kernel size of 4.

For fusion, all the priors including $f_{r+}^{u}(z)$, $f_{r-}^{u}(z)$, $f_{r+}^{b}(z)$, and $f_{r-}^{b}(z)$ and the target feature are fused to generate F_{I}' as shown in Fig. 7. Specifically, we concatenate the features and put the features into a reconstruction model [48] to obtain the images as in the MVIQE. When F_{It}', F_{Ix}', and F_{Iy}' are obtained along the temporal, x-axis, and y-axis directions, respectively, using the MPG modules, the intermediate frames are fused again to produce the final output F_{enh}. The procedure is presented in Fig. 4.

C. LOSS FUNCTION

For the loss function, we use Charbonnier loss function. Specifically, we first define a reconstruction loss function L_r to approximate the final output F_{eng} and the intermediate frames $F_{I'}$ in Fig. 4 to the ground-truth MVV frame F_{gt}

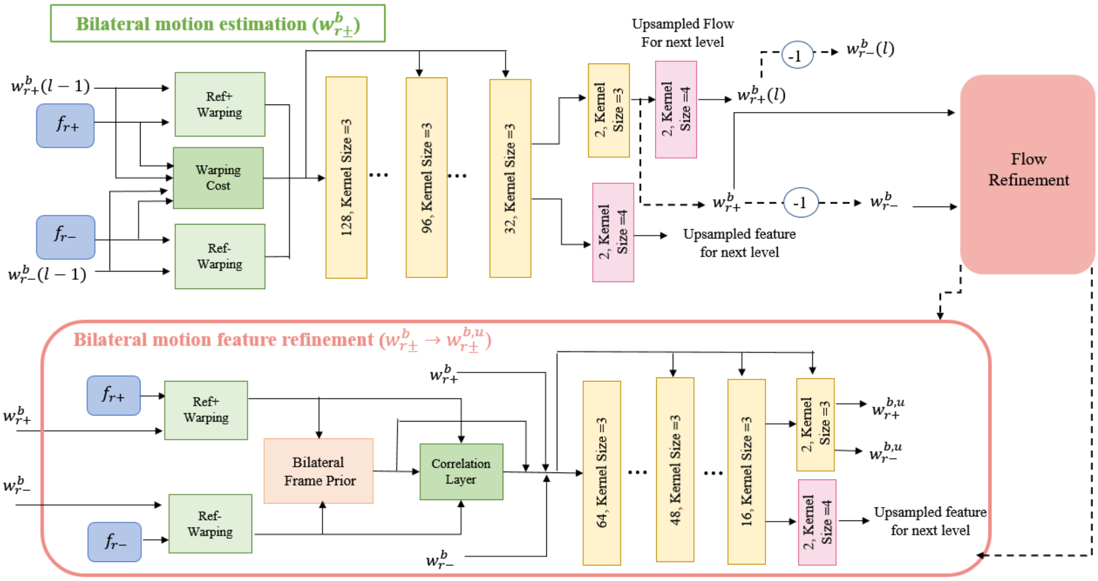


FIGURE 7. Illustration of bilateral motion estimation network blocks and refinement modules in the MPG model. The bilateral motion estimation blocks produce $w^b_{r\pm}$ and the warped features $f^b_{r\pm}(z)$. $w^b_{r\pm}$ are further refined with $f^b_{r\pm}(z)$ to produce $w^{b,u}_{r\pm}$ as presented in the orange blocks.

as follows:

$$L_r = \sum_{F_r \in \mathbf{R}} \sqrt{\left\| F_r - F_{gt} \right\|^2 + \rho^2}, \tag{10}$$

where $\mathbf{R} = \{F_{enh}, F_{Ix'}, F_{Iy'}, F_{It'}\}$, and ρ is set to 0.001.

We also use a warping loss function L_w to improve the accuracy of the warping. Using warping loss combined with reconstruction loss allows for the generation of accurate warping vectors. L_w is defined as follows:

$$L_w = \sum_{F_w \in \mathbf{W}} \sqrt{\left\| F_w - F_{gt} \right\|^2 + \rho^2}, \tag{11}$$

where F_w is the warped reference frame. $\mathbf{W}$ includes six warped reference frames using unidirectional vectors (i.e., F^u_{r-} and F^u_{r+} along t, x, and y directions) as a result of (3) and (4) and six warped reference frames using bilateral vectors (i.e., F^b_{r-} and F^b_{r+} along t, x, and y directions) as a result of (6) and (7). When the refinement module is used, w^b_{r+} and w^b_{r-} are replaced with $w^{b,u*}_{r+}$ and $w^{b,u*}_{r-}$, respectively, using (8) and (9) in an end-to-end learning. ρ is set to 0.001.

The final loss function is defined as in the following.

$$L_{total} = L_r + L_w. \tag{12}$$

IV. EXPERIMENTAL RESULT

A. TRAINING

1) DATASET

We use multi-view synthetic data to train the proposed network as in [49]. We used Unity software to generate virtual in-door and out-door scenes that displayed moving 3D foreground objects. In the scenes, there were several virtual cameras placed in 3D space, and the multi-view perspective videos were directly captured from the virtual cameras. The cameras are positioned along each of x, y, and z axes.

We have grouped multi-view videos with five views v_1, v_2, v_3, v_4, and v_5 in the same scene to train the proposed network. We define v_1 as the target view and select v_2, v_3, v_4, and v_5 as the reference views which are the source of spatial reference frames. The temporal reference frames for training are also located at v_1 but are sampled at the forward and backward time steps of the target frame.

We convert the synthetic training data into YUV video format and encode it with different quantization parameter (QP) using a MV-HEVC reference software. Since the reference frames are coded with relatively high quality than the target frame in 3D video coding standards, we encode reference frames with lower QP values than the QP value of the target. We utilized only Y frame in the training.

2) TRAINING DETAILS

The training dataset and the validation dataset are 1,200 frames and 400 frames respectively. For training, we extract 128×128 patches from the current frame and reference frame. The input patches are rotated either horizontally or vertically with a probability of 0.5. The batch size is 4. We train the network with a learning rate of 10^{-4}, a maximum epoch of 25,000, and use the Adam optimizer [50]. Our network was trained and implemented in pyTorch [51] and run on machines equipped with multiple RTX 2080 Ti Graphical Processing Units (GPUs).

B. TEST CONFIGURATION

To demonstrate the performance of the proposed network, we used MPEG-I MVV dataset for testing [52]. We used four different video sequences, consisting of "Kitchen", "Cadillac", and "Mirror", which are computer-generated (CG) data, and "Painter" as natural content (NC). All four

TABLE 1. Quantitative evaluation using PSNR values in MPEG-I test video sequences with HEVC multi-view extension.

Sequence	Tar QP	Ref QP	Target	CrossNet [18]	MVIQE [25]	TTSR [19]	MFQE2 [39]	Ours
Mirror	27	22	38.98	39.02	39.00	39.31	37.16	40.68
	32	27	36.25	36.27	36.26	36.41	35.74	38.26
	37	32	33.46	33.52	33.31	33.42	33.61	35.60
	42	37	30.69	30.69	30.31	30.45	30.99	32.85
Cadillac	27	22	42.27	42.38	42.36	43.17	40.80	43.73
	32	27	39.99	40.06	40.20	40.54	39.57	42.03
	37	32	37.31	37.53	37.85	37.60	37.58	39.85
	42	37	34.42	34.62	35.04	34.50	34.83	37.25
Kitchen	27	22	40.44	40.39	40.36	41.02	38.59	42.56
	32	27	37.66	37.66	37.58	37.83	37.05	39.89
	37	32	34.98	34.97	34.74	35.04	35.07	37.16
	42	37	32.39	32.33	31.87	32.31	32.77	34.33
Painter	27	22	40.04	40.20	40.10	40.91	39.33	40.74
	32	27	38.58	38.63	38.66	38.92	38.34	39.97
	37	32	36.68	36.76	36.89	36.88	36.77	38.51
	42	37	34.44	34.52	35.07	34.52	34.66	36.84
Average			36.79	36.85	36.85	37.05	36.43	38.77

TABLE 2. Quantitative evaluation using PSNR values in MPEG-I test video sequences with TMIV software.

Sequence	Tar QP	Ref QP	Target	CrossNet [18]	MVIQE [25]	TTSR [19]	MFQE2 [39]	Ours
Mirror	33	25	31.68	31.69	31.68	31.67	31.53	32.52
	42	33	29.45	29.48	29.46	29.47	29.55	30.82
	51	42	26.28	26.32	26.30	26.35	26.35	27.71
	61	51	23.34	23.36	23.38	23.35	23.37	24.10
Cadillac	31	21	33.42	33.43	33.41	33.44	33.35	33.78
	41	31	31.42	31.44	31.44	31.43	31.50	32.17
	51	41	28.53	28.59	28.88	28.61	28.66	29.76
	61	51	25.56	25.64	25.97	25.62	25.62	26.50
Kitchen	26	17	32.02	32.04	32.02	32.04	32.00	32.15
	33	26	31.59	31.62	31.54	31.61	31.64	31.91
	41	33	30.41	30.44	30.32	30.42	30.54	31.14
	49	41	28.31	28.36	28.25	28.46	28.49	29.59
Painter	32	21	37.97	37.96	37.98	37.93	36.72	38.81
	43	32	35.10	35.12	35.09	34.91	34.44	36.97
	51	43	31.65	31.69	32.02	31.81	31.26	33.45
	59	51	28.24	28.28	28.76	28.31	27.94	29.60
Average			30.31	30.34	30.41	30.34	30.19	31.31

sequences have a 2D camera array along the x-axis and y-axis.

For the HEVC experiments, the number of video frames is 48, 49, 49, and 60 for "Kitchen", "Cadillac", "Mirror", and "Painter", respectively. We used MIV common test conditions (CTC) specifications in [52] using 65 frames for TMIV experiments. TMIV CTCs specify the sequence-dependent QPs, denoted as RP1, RP2, RP3, and RP4. We followed the pre-defined QPs in the CTCs. All target and reference frames are cropped by 1024 × 1024.

C. QUANTITATIVE PERFORMANCE COMPARISONS

We show the quantitative performance of the proposed method denoted by "Ours." We calculate the PSNR values for quantitative comparisons. The results are evaluated with several state-of-the art studies, including CrossNet [18], MVIQE [25], TTSR [19], and MFQE2 [39]. We used a pre-trained model of MFQE2. The other tested methods have been trained with the same conditions in Sec. IV-A.

Table 1 shows the PSNR results with MPEG-I test video sequences, using HEVC multi-view extension software. The proposed method improves the PSNR values of the target video frames by approximately 2 dB and significantly outperforms the tested methods with different QPs. TTSR provided comparable results in the "Painter" sequence with the target QP (Tar QP) 27. However, the proposed method provides the best performance on the average. The performance improvements vary with the target QP values. When the videos include less quantization noise, e.g. Tar QP=27, the difference of the PSNR value in the "Painter" sequence was 0.7 dB between Ours and Target. On the other hand, the difference was approximately 2.4 dB in Tar QP=42. The performance of the proposed method was more reliable, when the videos had more severe noise. The priors have been helpful to enhance the video quality. We observed similar phenomenons in the other test sequences.

In Table 2, we have examined the performance of the PSNR values, when the target videos are coded with TMIV reference

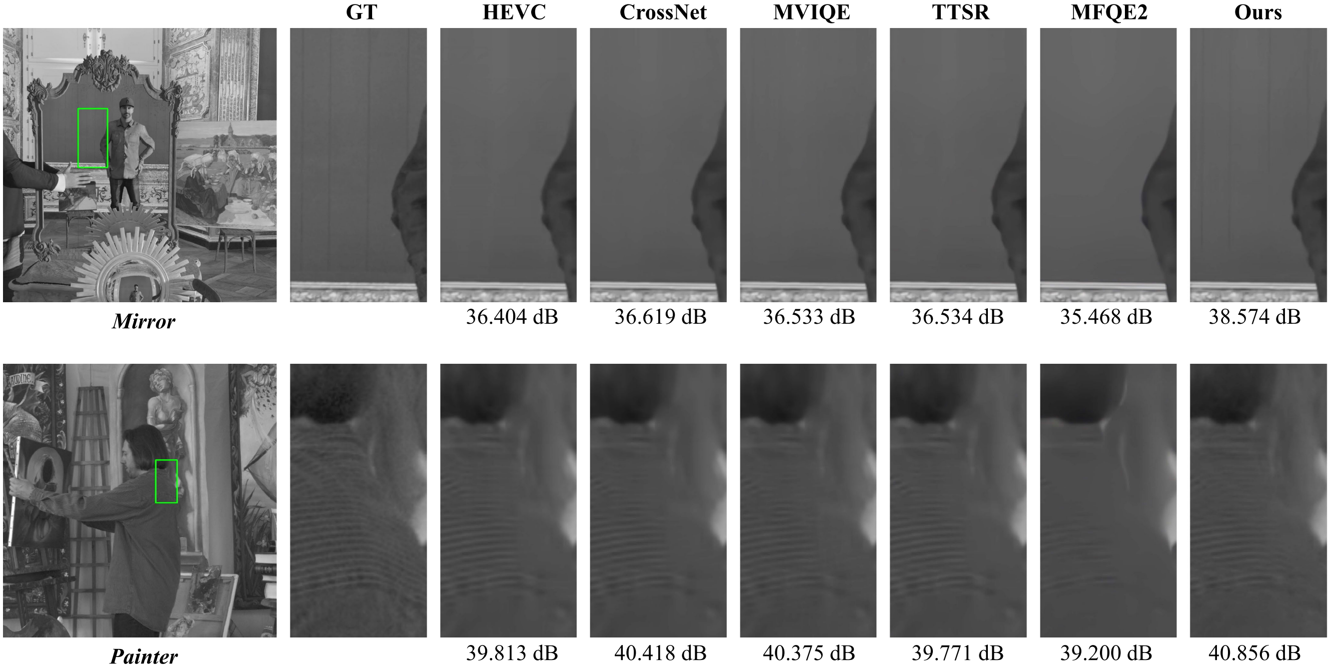


FIGURE 8. Qualitative comparison in MPEG-I test video sequences with HEVC software.

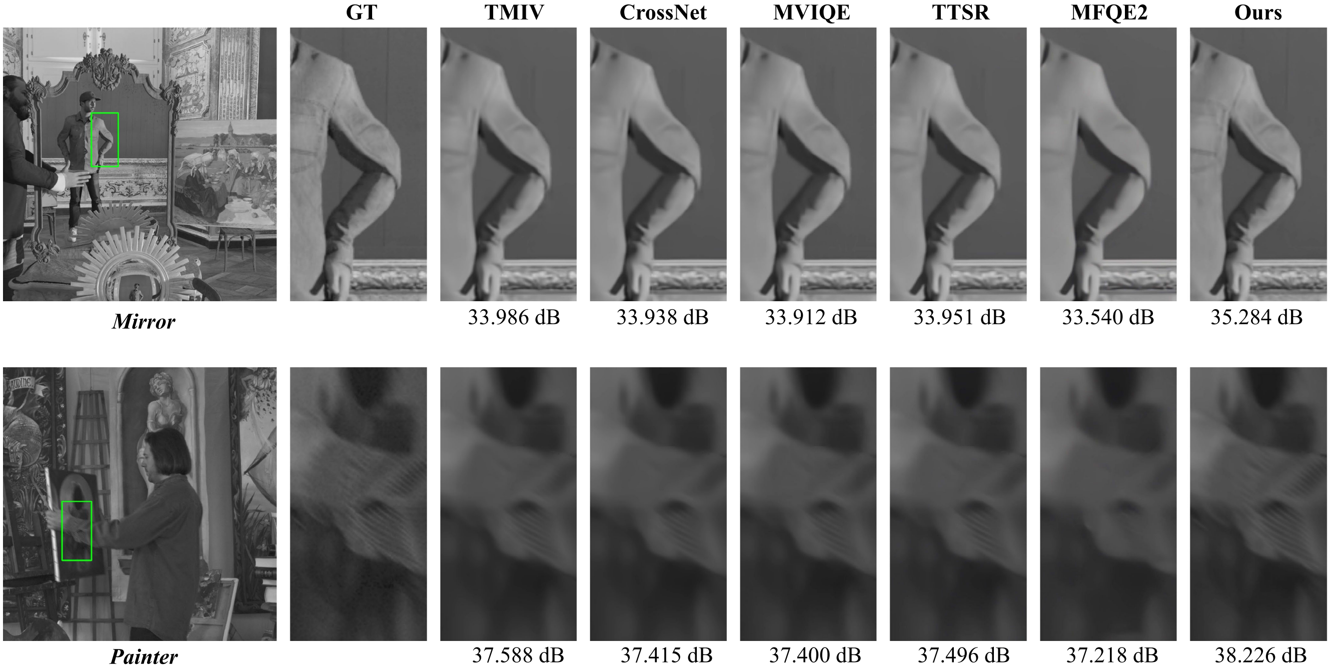


FIGURE 9. Qualitative comparison in MPEG-I test video sequences with TMIV software.

software. The PSNR values of the target video frames have been enhanced by approximately 1 dB. The performance improvements were relatively small as compared to the improvements in Table 1. In TMIV reference software, the target and the reference frames are reconstructed with atlases, which are synthesized using several coding tools and geometry information. Pixel pruning is employed to decrease redundant pixels in neighboring views through depth-based warping. The pruned pixels are arranged and packed into one or more atlases. Although these procedures were effective to minimize the pixel redundancies, it has been also observed that the reconstructed videos included not only quantization noise but also visible view synthesis artifacts. Synthesis artifacts cause a blurry effect across significant areas of the

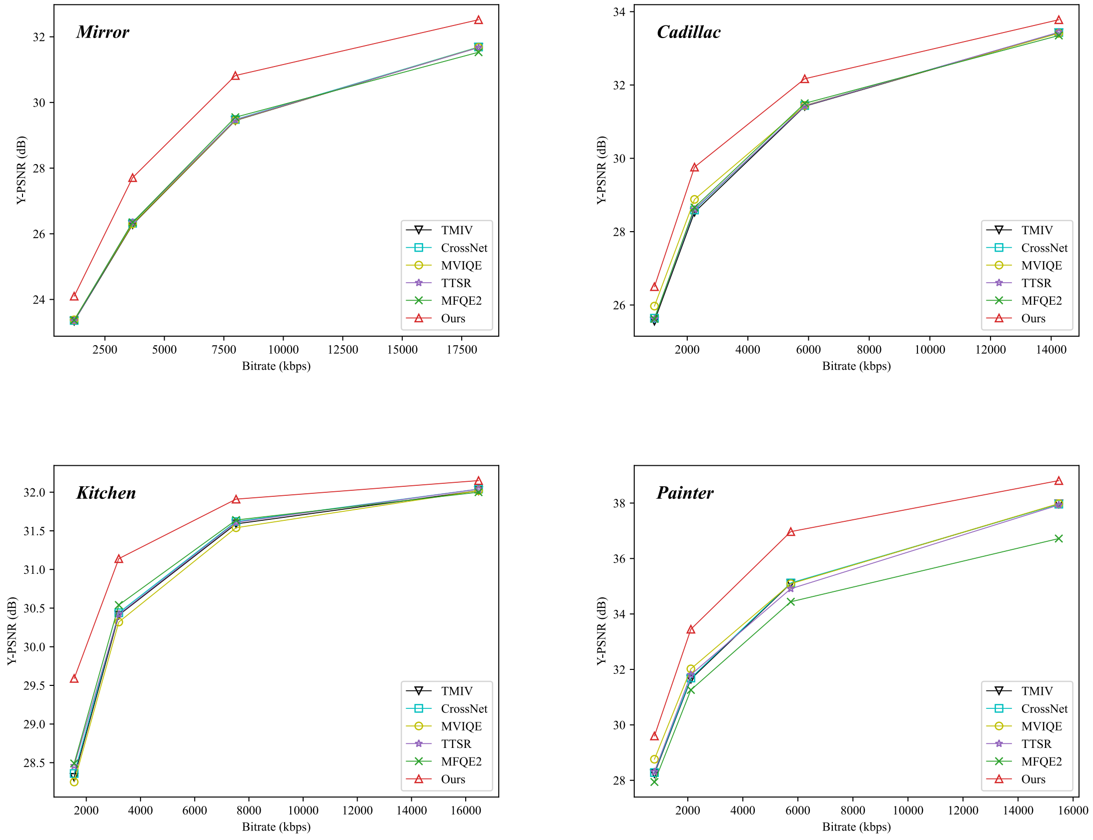


FIGURE 10. R-D curves in MPEG-I test video sequences with TMIV software.

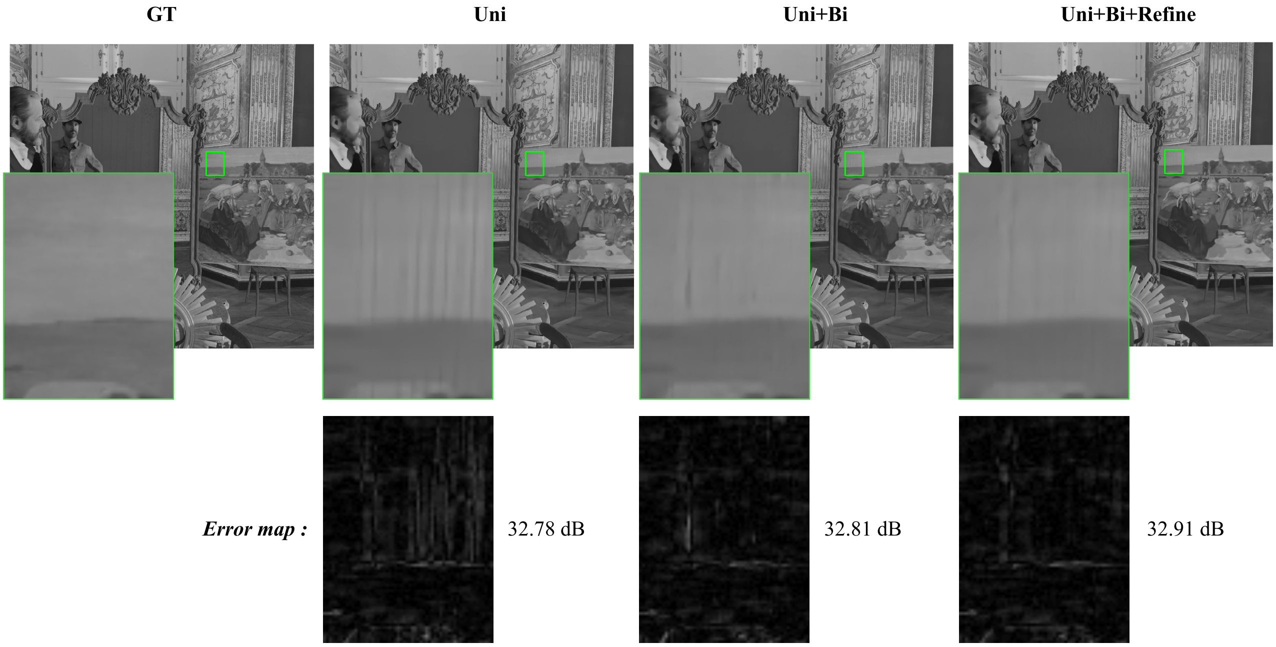


FIGURE 11. Qualitative comparison for ablation tests to examine various warping vectors.

target frames unlike quantization noise. In this situation, the accuracy of the previous motion prior with uni-prediction is compromised when attempting it to align it with the corresponding textures in the reference frames. For instance, the Crossnet suffered from the substantial degradation of the quality. However, the proposed motion priors can overcome this problem because they also try to match the reference-to-reference textures with bilateral motion priors. Our network

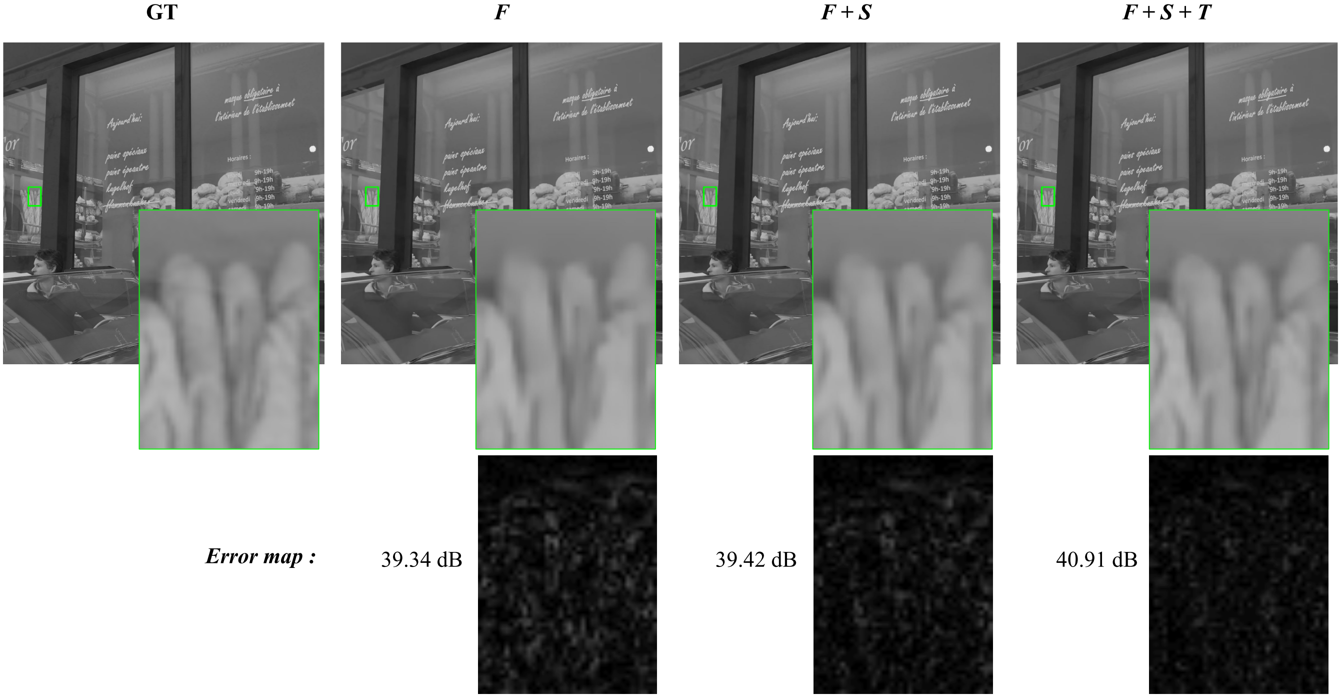


FIGURE 12. Qualitative comparison for ablation tests to examine the performance in fusing MPGs.

has been developed to incorporate various motion priors to be robust to quantization noise, but also the experimental results demonstrated that the proposed method can be effectively applied to TMIV reference software.

1) ABLATION TESTS

The proposed method uses various motion priors, including unidirectional warping vectors (U), bilateral warping vectors (B), and their refined vectors (R). We evaluate each contribution by enabling modules one-by-one. Table 3 presents the incremental performance in U, $U + B$, and $U + B + R$. As shown, the performance is improved with U approximately by 1.8 dB and increased with $U + B$ and $U + B + R$ approximately by 1.96 dB and 2.04 dB, respectively. The visual results are presented in Fig 11. Whereas unidirectional warping vectors used unreliable textures in the reference, the bilateral warping vectors and the refined vectors were able to refine the distortion.

We also conducted ablation tests to examine the performance in fusing MPGs. In Fig. 4, we used the priors (T) to a temporal direction and the priors (S) to spatial directions of both x and y axes in addition to F_{tar} denoted as F. The performance of the proposed method has been improved by 0.13 dB, 0.22 dB, and 2.04 dB, respectively, using F, S, and T. The visual results are exhibited in in Fig 12.

2) RATE-DISTORTION PERFORMANCE

We evaluate the R-D performance of the tested methods in Fig. 10. The BD-rate savings of the proposed method are −32.6%, −28.8%, −36.5%, and −41.0% for “Mirror”, “Cadillac”,“Mirror”,“Kitchen”, and “Painter” test sequences. As compared to the tested methods, the R-D curves demonstrate that the proposed method significantly improved the performance of the reconstructed frames in post-processing.

TABLE 3. Ablation tests of the MPG module. ***U***, ***B***, and ***R*** refer to unidirectional, bilateral, and refined warping vectors, respectively. The incremental PSNR values are presented to reveal each contribution within MPG. ***F***, ***S***, and ***T*** refer to ***F_{tar}***, spatial priors, and temporal priors, respectively, during fusion. Cadillac sequence in Tar QP=27 was used for the tests.

U	$U+B$	$U+B+R$	Δ PSNR
✓			Δ1.81 dB
✓	✓		Δ1.96 dB
✓	✓	✓	Δ2.04 dB
F	$F+S$	$F+S+T$	Δ PSNR
✓			Δ0.13 dB
✓	✓		Δ0.22 dB
✓	✓	✓	Δ2.04 dB

D. QUALITATIVE PERFORMANCE COMPARISONS

We compare the perceptual quality of tested methods using frame-by-frame visual comparisons. Fig. 8 and Fig. 9 show the visual comparisons with “Mirror” and “Painter” coded with the HEVC and TMIV software, respectively. It was demonstrated that the proposed method provided better visual quality in the video frames. For example, in Fig. 9, the texture of the clothe was blurred in the target frame, and it was difficult to search the corresponding textures. CrossNet, TTSR, and MFQE2 relies on the uni-directional motion priors, and they are difficult to improve the performance, when either a target or a reference have missed the relevant texture features. However, the bilateral motion priors use the reference-to-reference texture matching. Thus, the textures could be restored using the richer ones from the reference.

V. CONCLUSION

The paper addressed the challenge of reducing compression artifacts in MVV data coded by 3D video systems. These artifacts can negatively impact the quality of rendered 3D spaces. We proposed an efficient method to enhance the MVV quality by using spatial and temporal motion priors. Previous approaches employed motion priors for quality enhancement, but they were sensitive to noise and suffered from degraded warping accuracy. To tackle this, the paper introduced a new method that utilized bilateral motion priors, leveraging the structured geometry of the MVV camera system for motion coherence. This involved generating unidirectional and bilateral warping vectors for robust feature extraction from adjacent reference MVV frames. These features were refined to account for challenges such as occlusions. The effectiveness of the proposed method was demonstrated through experiments and comparison with existing techniques. The proposed method significantly improved the quality of reconstructed MVV frames in modern video coding standards such as the HEVC multi-view extension and the MPEG immersive video standard.

ACKNOWLEDGMENT

(Gyulee Jeon, Yeonjin Lee, and Jung-Kyung Lee are co-first authors.)

REFERENCES

[1] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the multiview and 3D extensions of high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 35–49, Jan. 2016.

[2] J. M. Boyce, R. Doré, A. Dziembowski, J. Fleureau, J. Jung, B. Kroon, B. Salahieh, V. K. M. Vadakital, and L. Yu, "MPEG immersive video coding standard," *Proc. IEEE*, vol. 109, no. 9, pp. 1521–1536, Sep. 2021.

[3] P. Garus, F. Henry, J. Jung, T. Maugey, and C. Guillemot, "Immersive video coding: Should geometry information be transmitted as depth maps?" *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 3250–3264, May 2022.

[4] H.-J. Kim, J.-W. Kang, and B.-U. Lee, "Super-resolution of multi-view ERP 360-degree images with two-stage disparity refinement," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 1283–1286.

[5] M. Xu, D. Niyato, J. Kang, Z. Xiong, C. Miao, and D. In Kim, "Wireless edge-empowered metaverse: A learning-based incentive mechanism for virtual reality," 2021, *arXiv:2111.03776*.

[6] Y. Chen, X. Zhao, L. Zhang, and J.-W. Kang, "Multiview and 3D video compression using neighboring block based disparity vectors," *IEEE Trans. Multimedia*, vol. 18, no. 4, pp. 576–589, Apr. 2016.

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

[8] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.

[9] J.-W. Kang, Y. Chen, L. Zhang, and M. Karczewicz, "Low complexity neighboring block based disparity vector derivation in 3D-HEVC," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Jun. 2014, pp. 1921–1924.

[10] L. Zhang, J. Kang, X. Zhao, Y. Chen, and R. Joshi, "Neighboring block based disparity vector derivation for 3D-AVC," in *Proc. Vis. Commun. Image Process. (VCIP)*, 2013, pp. 1–6.

[11] Y. Jin, M. Chen, T. Goodall, A. Patney, and A. C. Bovik, "Subjective and objective quality assessment of 2D and 3D foveated video compression in virtual reality," *IEEE Trans. Image Process.*, vol. 30, pp. 5905–5919, 2021.

[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] M. Ehrlich, L. Davis, S. N. Lim, and A. Shrivastava, "Quantization guided jpeg artifact correction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 293–309.

[14] Y. Kim, J. W. Soh, and N. I. Cho, "AGARNet: Adaptively gated JPEG compression artifacts removal network for a wide range quality factor," *IEEE Access*, vol. 8, pp. 20160–20170, 2020.

[15] A. Davy, T. Ehret, J.-M. Morel, P. Arias, and G. Facciolo, "A non-local CNN for video denoising," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2409–2413.

[16] M. Tassano, J. Delon, and T. Veit, "FastDVDnet: Towards real-time deep video denoising without flow estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1351–1360.

[17] L. Yu, L. Shen, H. Yang, L. Wang, and P. An, "Quality enhancement network via multi-reconstruction recursive residual learning for video coding," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 557–561, Apr. 2019.

[18] H. Zheng, M. Ji, H. Wang, Y. Liu, and L. Fang, "CrossNet: An end-to-end reference-based super resolution network using cross-scale warping," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 88–104.

[19] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.

[20] S. Zhou, Y.-H. Hu, and H. Jiang, "Multi-view image denoising using convolutional neural network," *Sensors*, vol. 19, no. 11, p. 2597, Jun. 2019.

[21] 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.

[22] M. Khamassi, M. Kaaniche, and A. Benazza-Benyahia, "Joint denoising of stereo images using 3D CNN," in *Proc. 10th Int. Symp. Signal, Image, Video Commun. (ISIVC)*, Apr. 2021, pp. 1–6.

[23] Z. Pan, W. Yu, J. Lei, N. Ling, and S. Kwong, "TSAN: Synthesized view quality enhancement via two-stream attention network for 3D-HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 345–358, Jan. 2022.

[24] B. Peng, R. Chang, Z. Pan, G. Li, N. Ling, and J. Lei, "Deep in-loop filtering via multi-domain correlation learning and partition constraint for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1911–1921, Apr. 2023.

[25] G.-L. Jeon, H.-J. Kim, E. Yeo, and J.-W. Kang, "CNN based multi-view image quality enhancement," in *Proc. 13th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2022, pp. 372–375.

[26] 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.

[27] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1461–1473, Nov. 2007.

[28] Z. Li, S. Niklaus, N. Snavely, and O. Wang, "Neural scene flow fields for space-time view synthesis of dynamic scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6494–6504.

[29] S. Niklaus and F. Liu, "Softmax splatting for video frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5436–5445.

[30] N. Kim and J.-W. Kang, "Dynamic motion estimation and evolution video prediction network," *IEEE Trans. Multimedia*, vol. 23, pp. 3986–3998, 2021.

[31] J. Park, K. Ko, C. Lee, and C.-S. Kim, "BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 109–125.

[32] 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.* Cham, Switzerland: Springer, 2017, pp. 28–39.

[33] S. Zhang, Z. Fan, N. Ling, and M. Jiang, "Recursive residual convolutional neural network-based in-loop filtering for intra frames," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1888–1900, Jul. 2020.

[34] J. Guo and H. Chao, "One-to-many network for visually pleasing compression artifacts reduction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4867–4876.

[35] L. Zhu, Y. Zhang, S. Wang, H. Yuan, S. Kwong, and H. H.-S. Ip, "Convolutional neural network-based synthesized view quality enhancement for 3D video coding," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5365–5377, Nov. 2018.

[36] S. Jammal, T. Tillo, and J. Xiao, "Multiview video quality enhancement without depth information," *Signal Process., Image Commun.*, vol. 75, pp. 22–31, Jul. 2019.

[37] X. He, Q. Liu, and Y. Yang, "MV-GNN: Multi-view graph neural network for compression artifacts reduction," *IEEE Trans. Image Process.*, vol. 29, pp. 6829–6840, 2020.

[38] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3892–3901.

[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.*, vol. 43, no. 3, pp. 949–963, Mar. 2021.

[40] 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)*, Sep. 2019, pp. 934–938.

[41] S. Chen, Q. Liu, and Y. Yang, "Adaptive multi-modality residual network for compression distorted multi-view depth video enhancement," *IEEE Access*, vol. 8, pp. 97072–97081, 2020.

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

[43] L. Yu, T. Tillo, J. Xiao, and M. Grangetto, "Convolutional neural network for intermediate view enhancement in multiview streaming," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 15–28, Jan. 2018.

[44] D. Mieloch, P. Garus, M. Milovanovic, J. Jung, J. Y. Jeong, S. L. Ravi, and B. Salahieh, "Overview and efficiency of decoder-side depth estimation in MPEG immersive video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6360–6374, Sep. 2022.

[45] B.-D. Choi, J.-W. Han, C.-S. Kim, and S.-J. Ko, "Motion-compensated frame interpolation using bilateral motion estimation and adaptive overlapped block motion compensation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 4, pp. 407–416, Apr. 2007.

[46] J. Jung and B. Kroon. *Common Test Conditions for MPEG Immersive Video*, Standard ISO/IEC JTC1/SC29/WG04 N, 2020.

[47] J. Park, C. Lee, and C.-S. Kim, "Asymmetric bilateral motion estimation for video frame interpolation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14519–14528.

[48] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1132–1140.

[49] H.-J. Kim, J.-W. Kang, and B.-U. Lee, "360° image reference-based super-resolution using latitude-aware convolution learned from synthetic to real," *IEEE Access*, vol. 9, pp. 155924–155935, 2021.

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[51] N. Ketkar, "Introduction to PyTorch," in *Deep Learning With Python*. Berlin, Germany: Springer, 2017, pp. 195–208.

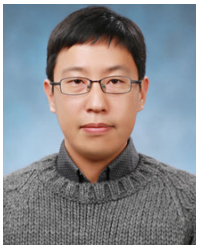
[52] J. J. E. A. Dziembowski and B. Kroon. *Common Test Conditions for MPEG Immersive Video*, Standard ISO/IEC JTC1/SC29/WG04 N 0342, Apr. 2023.

GYULEE JEON received the B.S. and M.S. degrees in electronics engineering from Ewha Womans University, Seoul, Republic of Korea, in 2018 and 2021, respectively. She is currently with LG Innotek. Her research interest includes image signal processing in a camera systems.

YEONJIN LEE received the B.S. degree in electronics engineering from Ewha Womans University, Seoul, Republic of Korea, in 2018, where she is currently pursuing the M.S. degree with the Department of Electronic and Electrical Engineering. Her research interest includes multi-view video quality enhancement.

JUNG-KYUNG LEE received the B.S. degree in electronics engineering from Ewha Womans University, Seoul, Republic of Korea, in 2018, where she is currently pursuing the Ph.D. degree with the Department of Electronic and Electrical Engineering. Her research interests include video compression and machine learning.

YONG-HWAN KIM was born in Jeju, South Korea, in 1972. He received the B.S. and M.S. degrees in electrical engineering and the Ph.D. degree in image engineering from Chung-Ang University, Seoul, South Korea, in 1996, 1998, and 2008, respectively. From 1999 to 2001, he was with Sungjin C&C, Seoul, where he optimized MPEG-1/2 video CODEC for DVR. Since 2001, he has been with KETI, Seongnam-si, South Korea, where he is currently a Chief Researcher with the Intelligent Image Processing Research Center. His current research interests include AV1, V-PCC, V-DMC, MIV video coding, and its optimized implementation.



JE-WON KANG (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering and computer science from Seoul National University, Seoul, South Korea, in 2006 and 2008, respectively, and the Ph.D. degree in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2012. He was a Senior Engineer with the Multimedia Research and Development and Standard Team, Qualcomm Technologies Inc., San Diego, CA, USA, from 2012 to 2014. He was a Visiting Researcher with the Nokia Research Center, Tampere University, Tampere, Finland, in 2011, and the Mitsubishi Electric Research Laboratories, Boston, MA, USA, in 2010. He has been an active Contributor to the recent international video coding standards in JCT-VC, including high-efficiency video coding (HEVC) standard and the extensions to multiview videos, 3-D videos, and screen content videos. He is currently an Associate Professor with Ewha Womans University, Seoul, where he is also the Head of the Information Coding and Processing Laboratory, Department of Electronics and Electrical Engineering. His current research interests include image and video processing and compression, computer vision, and machine learning.