IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 5 December 2022, accepted 31 December 2022, date of publication 9 January 2023, date of current version 18 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234917

RESEARCH ARTICLE

Saving Bits Using Multi-Sensor Collaboration

ZHE JI[1], HUI LAN[1], CHEOLKON JUNG[1], (Member, IEEE), DAN ZOU[2], AND MING LI[2]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Guangdong OPPO Mobile Telecommunications Corporation, Dongguan 523860, China

Corresponding author: Cheolkon Jung (zhengzk@xidian.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61872280 and Grant 62111540272.

ABSTRACT In this paper, we propose a new video coding method that saves bits using multi-sensor collaboration. Traditional video coding methods have saved bits by removing redundancy in videos. Recently, multiple types of sensors are being deployed to many solutions and multi-sensor data have significant advantages over single sensor data. The proposed method suggests a new way of video compression that saves bits using multi-sensor collaboration. We apply multi-sensor collaboration to the 3D video coding based on color and depth sensors. Based on the correlation between color and depth images, we design two networks CNN-US and CNN-QE in the proposed video coding method to achieve up-sampling and quality enhancement, respectively. The proposed method combines CNN-US and CNN-QE with 3D-HEVC to save bits using multi-sensor collaboration. Compared with 3D-HEVC anchor, the proposed method achieves average 5.9%, 66.8%, and 71.0% BD-rate reductions for sampling factors 1, 2, and 4 on the depth videos of 3D-HEVC test dataset, respectively.

INDEX TERMS 3D-HEVC, convolutional neural network, multi-sensor collaboration, redundancy, video coding.

I. INTRODUCTION

In recent years, the storage and transmission of video data have become more and more common, and a huge amount of video data have been produced persistently. Thus, the effective compression of video data is increasingly important. The video coding technology has made meaningful contributions to the compression of video data. The earliest research on video compression can be traced back to 1929 when inter-frame compression was first proposed. After years of research and development, mature video compression codec standards have gradually formed, such as MPEG-2 [1], MPEG-4 [2], and HEVC [3], [4]. MPEG-2 provides a wide range of compression rates to adapt to different picture quality, storage capacity and bandwidth requirements. However, the high-definition videos need higher compression efficiency, which has a limit by MPEG-2. MPEG-4 compresses and transmits video data through extremely narrow bandwidth and object-based coding to obtain the best image quality with the least amount of data. Compared with MPEG-2, MPEG-4 is suitable for interactive video services and remote surveillance. The High Efficiency Video Coding (HEVC) standard is based on the MPEG-4 framework and improves some modules such as inter-frame prediction, intra-frame prediction and in-loop filter. Under the same image quality condition, data compression rate of HEVC is 1.5 times higher than MPEG-4. The latest Versatile Video Coding (VVC) standard [5] was officially released in 2020, which represents the most advanced video coding technology at present. VVC is based on the HEVC coding framework and has further improved lots of modules. Otherwise, VVC has upgraded the encoding structure with multiple options such as concurrent processing of encoder and decoder. Compared with HEVC, VVC achieves nearly 50% bitrate reduction under the same perceptual quality. VVC encoding complexity is 10 times that of HEVC, while VVC decoding complexity is about 1.5 times that of HEVC. In recent years, 3D videos have received much attention due to the demands for virtual reality. Plenty of scenes adopt depth image-based rendering (DIBR) to generate a set of dense views, which needs high quality depth images. Therefore, 3D-HEVC [6] is investigated by JCT-3V as a 3D video coding standard [7]. 3D-HEVC is an extension on the basis of HEVC, which efficiently compresses multi-views and their corresponding

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

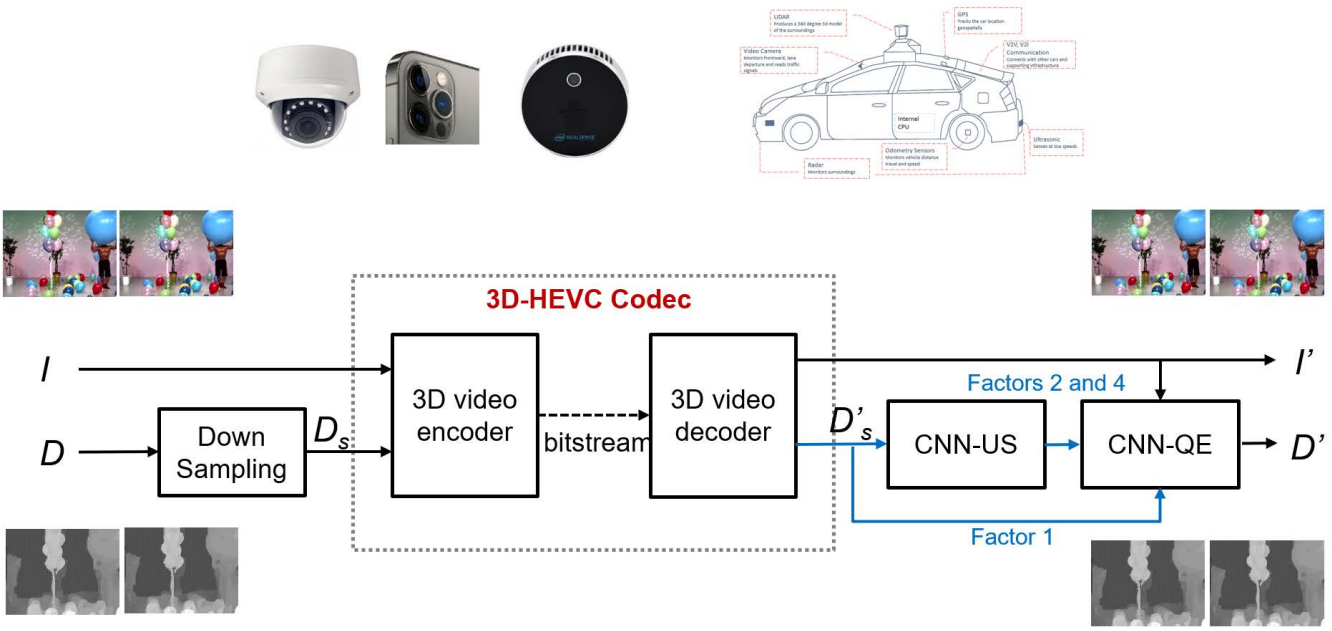

FIGURE 1. Entire framework of the proposed video coding method based on multi-sensor collaboration. The proposed method combines two networks CNN-US and CNN-QE with 3D-HEVC to save bits using multi-sensor collaboration. CNN-US is used to achieve up-sampling on the compressed depth video frames for sampling factors 2 and 4, while CNN-QE is used to achieve quality enhancement on the depth video frames based on the correlation between color and depth for all sampling factors, i.e. 1, 2, and 4.

depth data. 3D-HEVC includes all the key technologies of HEVC and employs new compression technologies that extract the unique characteristics of depth images and utilize the dependencies between multiple views as well as between texture and depth. Hence, 3D-HEVC has more advantages in the consumer applications that require video texture and depth. Compared with HEVC, 3D-HEVC specifically adapts to the properties of depth images, which satisfies the urgent need for depth image coding.

With the advent of deep learning, many methods based on deep neural networks have been proposed to enhance the coding efficiency of 3D-HEVC. Li et al. [8] proposed self-learning residual model-based fast coding unit (CU) size decision in the intra-coding of both texture views and depth images that utilized residual signal as the feature of CU to learn the features of the encoded coding tree unit (CTU). They achieved reduction of encoding time by the fast CU size decision. Zhang et al. [9] adopted a method of detecting the smooth area and texture direction in the depth image to reduce the number of intra-modes while decreasing the complexity and time cost. These methods were dedicated to modifying the internal modules of 3D-HEVC for performance improvement. With the recent advances in the sensor technology, especially the popularization of multi-sensory data, there is a new opportunity to reform and elevate the coding efficiency using **multi-sensor collaboration**. However, traditional video codecs, including 3D-HEVC, save bits by removing redundancy, and do not take multi-sensor collaboration into consideration to save bits. In addition to the redundancy removal, multi-sensor collaboration of color and depth images can remarkably contribute to improving the coding efficiency. 3D-HEVC achieves depth image coding but does not consider multi-sensor collaboration between color and depth images. Moreover, if quantization parameter (QP) is large, there would be obvious blocky artifacts in the decoded results. Although most of existing methods based on deep learning achieve speed-up of the prediction mode decision for coding unit/prediction unit (CU/PU), they are not robust to blocky artifacts under a large QP.

In this paper, we propose a new video coding method that can save bits using **multi-sensor collaboration**. We apply multi-sensor collaboration to the 3D video coding based on color and depth videos. Inspired by [25], we build two networks CNN-US and CNN-QE for the proposed method: CNN-US is for up-sampling of the depth videos in sampling factors 2 and 4, while CNN-QE is for quality enhancement of the depth videos based on the correlation between color and depth in all sampling factors 1, 2, and 4. First, we down-sample the depth video frames in sampling factors 2 and 4. Then, we utilize 3D-HEVC codec to encode and decode the input color and depth videos. Next, we adopt CNN-US to achieve up-sampling on the decoded depth video frames in sampling factors 2 and 4. Finally, based on the correlation between color and depth, we use CNN-QE to achieve quality enhancement on the depth video frames in all sampling factors, i.e. 1, 2 and 4. Through experiments, we found that down-sampling methods have little effect on the performance and thus we choose uniform sampling for down-sampling. Fig. 1 illustrates the proposed video coding method based on multi-sensor collaboration with consumer applications.

Compared with existing methods, main contributions of this paper are summarized as follows:

- We propose a new video coding method that saves bits using multi-sensor collaboration. We apply multi-sensor collaboration to 3D video coding based on color and depth videos, and use 3D-HEVC codec as baseline for the proposed method.
- We build two networks CNN-US and CNN-QE for color guided depth super-resolution (SR). CNN-US is used for depth up-sampling, while CNN-QE is for depth quality enhancement based on multi-sensor collaboration (color and depth). The proposed method considers three sampling factors 1, 2 and 4 based on CNN-US and CNN-QE.
- We verify the effectiveness of the proposed method for video compression in comparison with 3D-HEVC anchor. Compared with 3D-HEVC anchor, the proposed method achieves average 5.9%, 66.8%, and 71.0% BD-rate reductions for sampling factors 1, 2, and 4 on the depth videos of 3D-HEVC test dataset, respectively.

The rest of this paper is organized as follows. In Section II, we explain the advantage of multi-sensor collaboration and some relevant methods. Section III describes the proposed method based on 3D-HEVC codec, while Section IV provides visual comparison and quantitative measurements. Conclusions are made in Section V.

II. RELATED WORK

A. MULTI-SENSOR COLLABORATION

Accompanied by the continuous improvement of the sensor technology, various sensors such as depth, infrared (IR) and near-infrared (NIR) sensors have been widely utilized in recent years. Multi-sensory data are popular and being applied to many consumer electronics such as smartphones, self-driving cars and video surveillance. Since each type of sensors has its own characteristics, multi-sensory data are complementary. Thus, many outstanding achievements have been resulted in image super-resolution (SR), image fusion and object detection based on multi-sensor collaboration. In practice, multi-sensor collaboration is very similar to the cognition process of human brains. Human decision is made by analyzing various information obtained by sensory organs. Similar to this, multi-sensory data have significant advantages over single sensor data, which overcome the limitation of single modal data. Thus, multi-sensor collaboration has been widely applied to many kinds of computer vision tasks such as quality enhancement, scene reconstruction and target detection. Jiang et al. [10] proposed a deep edge guided depth SR method that included an edge prediction module and an SR module. The edge prediction module utilized hierarchical representation of color and depth images to produce accurate edge maps, which can promote the performance of SR module. Huang et al. [11] proposed a sparsity-invariant multi-scale encoder-decoder network (HMS-Net) for depth completion to handle sparse inputs and feature maps. They incorporated color information with depth information obtained by LIDAR camera to improve the performance in depth completion. Duan and Jung [12] proposed joint disparity estimation and pseudo near infrared (NIR) generation from cross spectral image pairs. They adopted difference map operator (DMO) and non-local blocks (NLB) to bridge the spectral gap between Y channel and NIR image. Chen et al. [13] proposed a sensor fusion framework that took both LIDAR point data and color image as input and predicted 3D bounding boxes for object detection in the autonomous driving environment. Hughes et al. [14] proposed a pseudo-siamese convolutional neural network (CNN) architecture to solve the task of identifying corresponding patches in very-high-resolution (VHR) optical and synthetic aperture radar (SAR) remote sensing imagery. These methods make full use of advantages from multiple sensors for computer vision tasks. Lan et al. [15] proposed a multi-sensor collaboration network for video compression based on wavelet decomposition, called MSCN. MSCN first combined multi-sensor collaboration with video compression.

B. 3D-HEVC

Video coding standards aims at removing redundancy in videos and saving bits, and are extended to supporting the representation of multiview videos and multiview plus depth formats. 3D-HEVC, as an 3D extension of HEVC, is targeted at a coded representation consisting of multiple views and associated depth images, generating additional intermediate views in advanced 3D displays. Compared with HEVC, additional bit rate reduction in 3D-HEVC is achieved by specifying new block-level video coding tools, which explicitly exploit statistical dependencies between texture and depth, and specifically adapt to the depth properties. In recent years, MPEG Immersive Video (MIV) standard [16] has been proposed. The draft MIV standard provides support for viewing immersive volumetric content captured by multiple cameras with six degrees of freedom (6DoF) within a viewing space determined by the camera arrangement. In the Test Model for Immersive Video (TMIV), multiple texture and geometry views are coded as atlases of patches using a legacy 2-D video codec, while optimizing for bit rate, pixel rate, and quality. The MIV standard enables a high-fidelity immersive experience through playback of camera-captured 3-D scenes with 6DoF of viewer position and orientation. It supports such consumer applications with affordable coded pixel rate and higher coding efficiency, especially for source content with high-quality depth information.

C. DEPTH IMAGE SUPER-RESOLUTION

Up to now, depth image SR works are divided into two categories: traditional approach and deep learning approach. Traditional methods are more flexible, while deep learning methods are good at obtaining the complex mapping functions from a large scale dataset. Traditional depth SR methods are further divided into three categories: learning-based methods, filtering-based methods and regularization-based methods. The core problem of learning-based methods is to

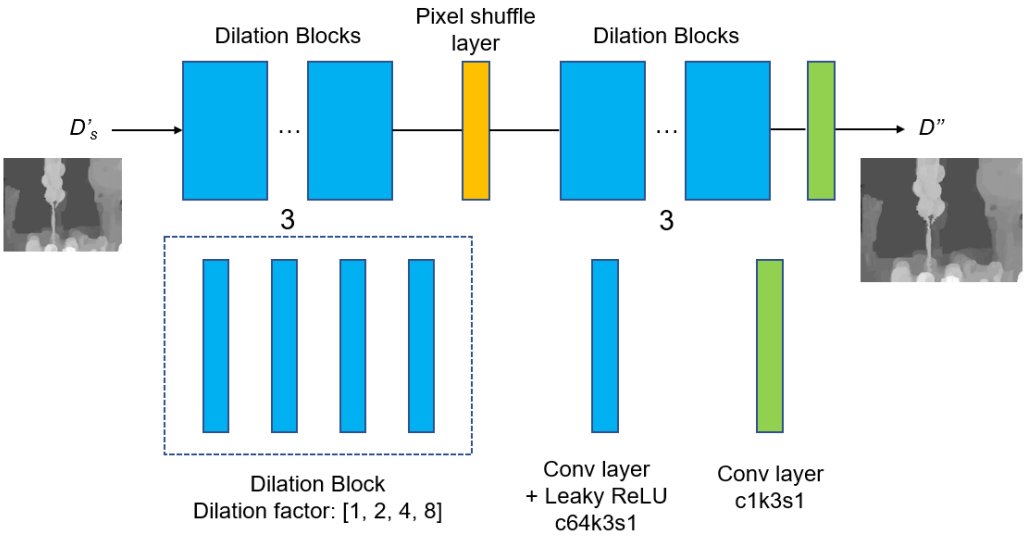

FIGURE 2. Network architecture of CNN-US. There are 3 dilation blocks before and after pixle shuffle layer. Every dilation block consists of 4 dilated convolution layers followed by Leaky ReLU layers. c64k3s1 indicates a feature map number of 64, a kernel size of 3 and a stride of 1. D'_S represents the decoded LR depth image and D'' represents the output upsampled by CNN-US.

obtain a sparse representation of depth images by designing dictionaries. Ferstl et al. [17] learned a dictionary of edge priors from an external database of high resolution (HR) and low resolution (LR) examples, which can be used in variational depth SR as an anisotropic guidance. Since global dictionaries can not adapt to local features of depth images well, Mandal et al. [18] proposed an edge preserving constraint and a pyramidal reconstruction strategy, which could preserve the discontinuity appeared in the depth image and deal with a higher upsampling factor. Filtering-based methods achieved depth SR via local filters, which usually relied on guidance maps. The representative work is joint bilateral fileter [19], which calculated the filter parameters using the RGB-D pairs for depth SR. Lo et al. [20] presented a joint trilateral filtering (JTF) algorithm for depth image SR, which extracted spatial and range information of local pixels and integrated local gradient information of the depth image. The regularization-based methods adopted regularization terms to make the depth SR problem well constrained. Liu et al. [21] proposed a robust optimization framework for color guided depth image restoration, which performed well in suppressing texture copy artifacts and preserved sharp depth discontinuities than the previous weighting schemes.

The application of convolution neural network (CNN) has greatly improved the performance of depth SR, which benefits from advanced network architecture, effective loss functions and massive data. Ye et al. [22] proposed an end-to-end deep controllable slicing network to realize region-level depth recovery and high generalization ability for the task of depth SR, which contains a scale-controllable module and a depth slicing module for realizing the fine-grained control of depth restoration with arbitrary magnification and using depth image features with different depth ranges. In CNN-based methods, color image is often adopted as supplementary information to improve reconstruction accuracy. In addition, these methods based on RGB-D pairs need extra operations to prevent texture artifacts. Jiang et al. [23] proposed to predict depth edges via fusing deep features extracted from two kinds of images in different scales without directly utilizing color images. They constructed a disentangling cascaded SR network to achieve depth image SR by fusing depth edge map and LR depth image. Deng et al. [24] designed a novel CNN to solve the general multi-modal image restoration (MIR) and multi-modal image fusion (MIF) problems based on a multi-modal convolutional sparse coding (MCSC) model.

Since multi-sensory data are complementary, e.g. color and depth, we propose a new video coding method that combines multi-sensor collaboration with video compression to save bits in this work.

III. PROPOSED METHOD

As illustrated in Fig. 1, the proposed method combines two networks CNN-US and CNN-QE with 3D-HEVC to save bits using multi-sensor collaboration. CNN-US is used to achieve up-sampling on the compressed depth video frames for sampling factors 2 and 4, while CNN-QE is used to achieve quality enhancement on the depth video frames based on the correlation between color and depth for all sampling factors, i.e. 1, 2, and 4.

A. CNN-US

CNN-US is proposed to achieve depth image super-resolution, which can be used to the case of sampling factors 2 and 4 in our framework. The network architecture of CNN-US is shown in Fig. 2. Dilated convolution [26] can increase the receptive field while keeping the number of parameters unchanged, which achieves that each convolution

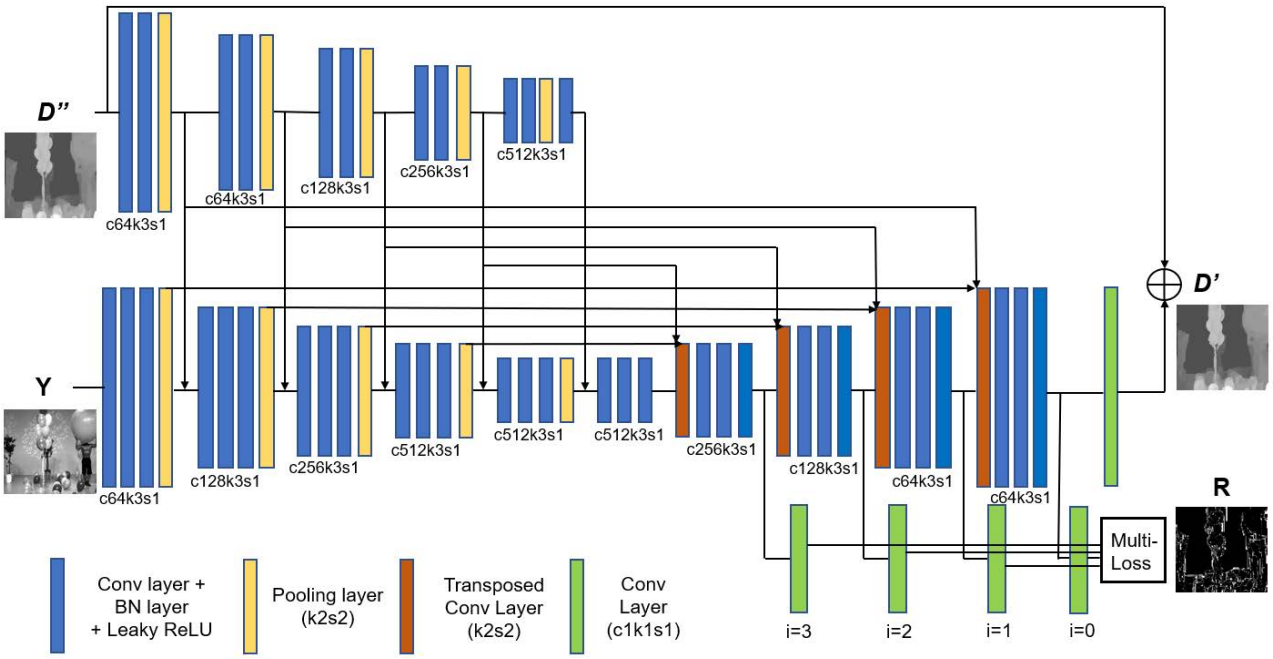

FIGURE 3. Network architecture of CNN-QE. CNN-QE is based on U-Net framework and residual learning, which utilizes the depth image as guidance to conduct the Y channel of the corresponding RGB image and generate its residual map. Meanwhile, CNN-QE performs intermediate prediction of each upsampled block output as multi-scale loss function. c64k3s1 indicates a feature map number of 64, a kernel size of 3 and a stride of 1, while i indicates different scales. D'' represents the output of CNN-US, Y represents the Y-channel of the corresponding compressed RGB image, D' represents the output of CNN-QE and R represents the difference between the final output and the ground truth of the depth image.

output contains rich context information, and ensure that the size of the output feature map remains constant. Therefore, dilated convolution can well avoid the loss of internal data structure and spatial hierarchical information caused by the upsampling layer and the pooling layer, and reconstruct the information of tiny objects. The pixel shuffle layer [27] converts low-resolution (LR) feature maps to high-resolution ones (HR) through convolution and multi-channel recombination, which can effectively avoid the artifacts during up-sampling by convolution and interpolation.

CNN-US utilizes pixel shuffle layer as up-sampling operation and dilation blocks to better capture global information of images. Each block is composed of 4 dilated convolution layers followed by the Leaky ReLU layers. The input of CNN-US is the frames of compressed low-resolution depth video, while the output of CNN-US is the frames of high-resolution depth video. We adopt $L2$-loss as the loss function of CNN-US, which is defined as follows:

$$L_{US} = \left\| D'' - GT \right\|_2 \tag{1}$$

where D'' represents the output of CNN-US and GT represents the corresponding ground truth of the depth image.

B. CNN-QE

CNN-QE is designed to achieve quality enhancment on depth images based on multi-sensor collaboration. Guo et al. [28] proposed a depth super-resolution method which infers a HR depth image from its LR version by hierarchical features driven residual learning. The method achieves depth image enhancement by obtaining a residual map corresponding to the up-sampled depth image via a convolutional neural network. Inspired by this idea, we designed CNN-QE to implement enhancement operation on the depth image by residual learning, which can enhance the high frequency component of depth video frames, to apply to sampling factors 1, 2 and 4 in our framework. The overview of the proposed network architecture and parameter settings is shown in Fig. 3. Different from previous depth super-resolution methods like [28] that extract hierarchical intensity features from color images to transfer useful structure to the final HR depth images, our proposed framework utlizes the structure information of depth images as guidance to assist the Y channel of the corresponding color images to reconstruct residual maps as shown in Fig. 3. CNN-QE uses the fixed-size convolution kernel to extract different levels of depth features, which can make full use of the edge information in the depth image and eliminate a large number of detailed textures in color images. In addition, the proposed CNN-QE is based on U-Net [29] framework. Skip connection operation is a direct connection between nodes of different layers in U-Net framework by skipping one or more layers of nonlinear processing. As one of the algorithms that utilize multi-scale features to solve problems, skip connection can alleviate gradient disappearance and achieve feature enhancement. Based on U-Net framework, CNN-QE can realize feature reuse and ensure maximum information flow between layers. Meanwhile, inspired by [30], we generated intermediate predictions of each upsampled block output and put them into the loss function, which can minimize the

difference between the reconstructed residual map and the corresponding ground truth. In addition, we introduced the difference between the final output and the ground truth of the depth image as a part of loss function to improve enhancement performance. In our experiments, L_2-loss is good enough to get better results in CNN-QE. The loss function of CNN-QE is formulated as:

$$L_{QE} = \sum_{i=0}^{3} 2^{-i} * \|R_i - m_i\|_2 + \left\|D'_s - GT\right\|_2 \quad (2)$$

where m_i represents multi-scale feature map and R_i represents the residual map of the corresponding size, m_0 represents the output of CNN-QE, R_0 represents the ground truth of the residual map, D' represents the input of CNN-QE and GT represents the ground truth of the depth image.

Based on the characteristics of multi-sensor collaboration and residual learning, CNN-QE with U-Net framework as backbone and L_{QE} as loss function, is designed to achieve depth image enhancement. The input of CNN-QE is decoded depth and color video frame or the output of CNN-US, and the output is the enhanced depth video frame.

C. MODEL SELECTION STRATEGY

In our experiements, we found that in sampling factor 1, i.e. the same size of the input color and depth videos, due to the difference of residual maps under different quantization parameters (QPs), the effect of a single training dataset is poor for the recovery of high frequency component of depth videos. Meanwhile, we found that the performance can be improved by a compressed training dataset whose compression degree, i.e. QP, is smaller than the compression degree of the test sequences. Thus, we use the compressed dataset for training whose QP is slightly lower than that of the test sequences. The loss of the data is similar between the training and testing sets, which is more conducive to the image reconstruction. Therefore, we use a model selection strategy for sampling factor 1 in the proposed method as follows:

1) For QP34, train CNN-QE with uncompressed training data.

2) For QP39 and QP42, train CNN-QE with training data compressed by HM16.16 in QP34.

3) For QP45, train CNN-QE with training data compressed by HM16.16 in QP39.

The setting of QPs is based on 3D-HEVC common test condition (CTC) [7].

IV. EXPERIMENTAL RESULTS

Compared with 3D-HEVC anchor, we perform visual comparison and quantitative measurements on 7 test sequences in 3D-HEVC dataset [7]. To consider the size mismatch between color and depth frames, the proposed method is implemented on 3D-HEVC codec by encoding and decoding color and depth videos separately.

A. NETWORK TRAINING AND IMPLEMENTATION

For sampling factors 2 and 4, we utilize the same training datasets with [28], namely 58 RGB-D images from MPI Sintel depth dataset [31] and 34 RGB-D images from Middlebury dataset [32]. To increase the amount of training data, we augment data with flipping and rotation [28]. In the training phase, the depth images are cropped to 128 × 128 image patches by random sampling, thus reducing the training time. Finally, the augmented training data have roughly 170,000 image patches. To synthesize LR depth images, we downsample each full-resolution image patch by uniform sampling with the scaling factors. For sampling factor 1, it is required to use the training data under different QPs to achieve depth video enhancement. Therefore, we use DIML indoor training dataset [33] that contains 1500 RGB-D images and generate the compressed training data in QPs 39, 42 and 45. We also perform the same data augmentation in sampling factors 2 and 4. Since the training datasets are image pairs, all of the test sequences are compressed by HEVC reference software, HM16.16, under All Intra (AI) configuration. The setting of QPs is based on 3D-HEVC CTC, where the QPs of the depth video are 34, 39, 42, 45 and the QPs of the corresponding color video are 25, 30, 35, 40. These test sequences covers different resolutions and scene conditions to verify the performance of the proposed method framework. During the training phase in CNN-US and CNN-QE, we use a batch-mode learning method with a batch size of 64 with ADAM optimizer for network optimization. The number of epochs is set to 100 for CNN-US and 50 for CNN-QE. The learning rate is set to $1e^{-4}$ for CNN-US and CNN-QE under different sampling factors. For training, we use PyTorch framework on a PC with one Tesla V100 GPU.

B. VISUAL COMPARISON

We evaluate the proposed method on 7 test sequences that are provided by 3D-HEVC CTC [7]. The 7 test sequences are composed of two groups according to size: one with size 1024 × 768 - *Kendo*, *Balloons* and *Newspaper*, and the other with size 1920 × 1088 - *Poznan_Hall2*, *Poznan_Street*, *Undo_Dancer* and *GT_Fly*. Since each group shows similar performance, we select one test sequence for visual comparison on each group: *Kendo* and *Undo_Dancer*. Meanwhile, in our experiments, we have found that HM can not encode the second group of test sequences whose size is 1920 × 1088 under sampling factor 2. The situation is due to that HM codec is unable to divide proper Coding Unit (CU) for inappropriate video size. Therefore, we have implemented sampling factors 1 and 4 on the second group.

The visual comparison results on *Kendo* and *Undo_Dancer* sequences are shown in Fig. 4 and Fig. 5, respectively. The results of the first row show that 3D-HEVC anchor occurs obvious blocky artifacts that expand with the increase of QP and edge information of depth images gradually blurs. The second row shows the results by the proposed method under sampling factor 1. Compared with 3D-HEVC anchor, the edge information has been enhanced to a certain extent.

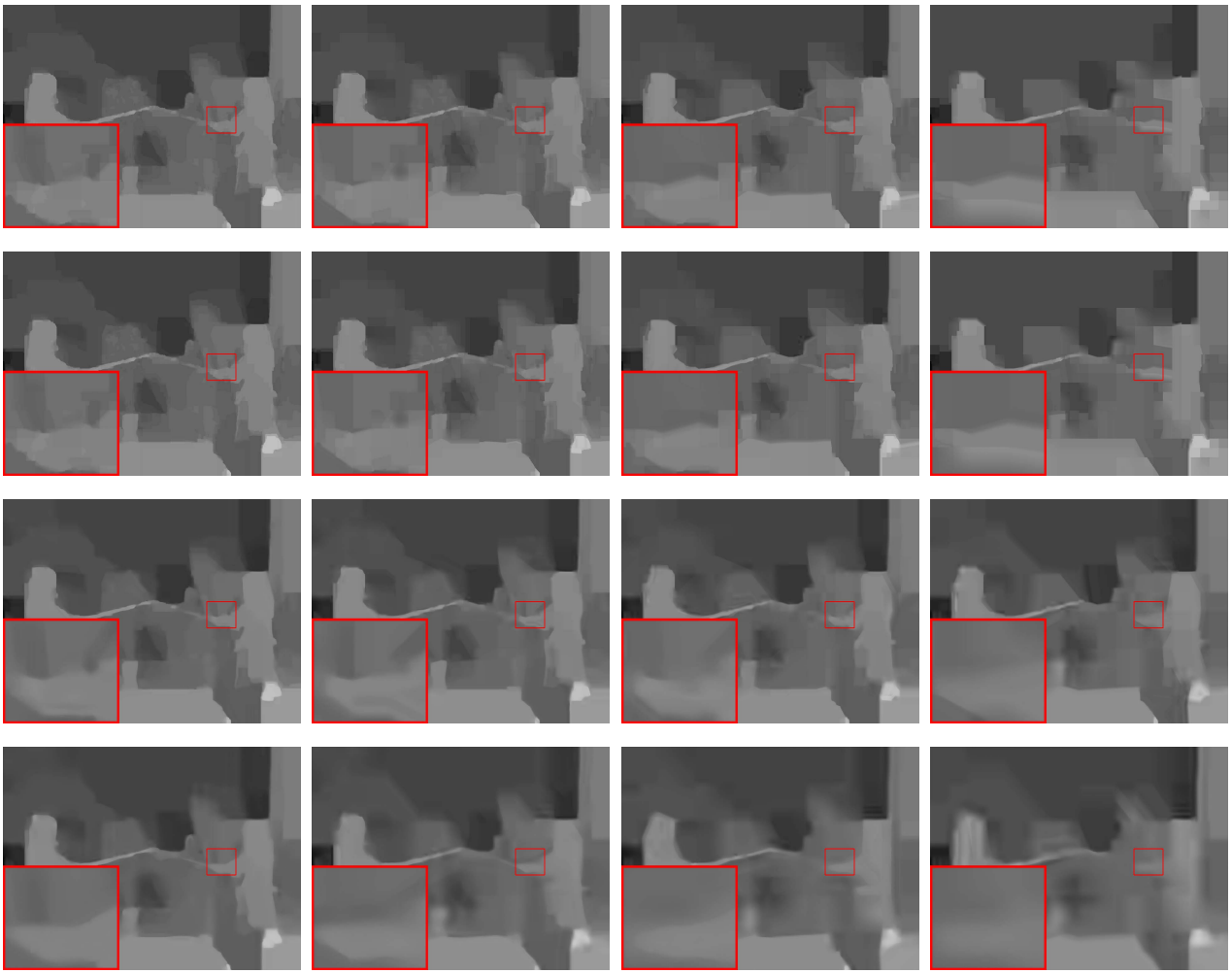

FIGURE 4. Visual comparison on *Kendo* sequence. QP represented by four columns are 34, 39, 42 and 45 in turn. The first row represents 3D-HEVC anchor with apparent blocky artifacts. The second row, the third row and the last row represent the experimental results under sampling factors 1, 2 and 4, respectively. The results indicate that the proposed method can eliminate blocky artifacts, but with local distortion in some cases.

TABLE 1. BD rates for the first view of depth videos in comparison with 3D-HEVC anchor on 7 test sequences. NR: Not reported.

	Kendo	*Balloons*	*Newspaper*	*GT_Fly*	*Poznan_Hall2*	*Poznan_Street*	*Undo_Dancer*
Sampling	BD rate	BD rate	BD rate	BD rate	BD rate	BD rate	BD rate
Factor 1	-6.4%	-5.2%	-1.3%	-11.1%	-1.3%	-10.5%	-5.5%
Factor 2	-61.2%	-67.7%	-71.6%	NR	NR	NR	NR
Factor 4	-71.0%	-68.2%	-70.7%	-75.9%	-66.1%	-80.1%	-65.1%

The third row shows the results under sampling factor 2. As QP increases, the proposed method causes edge blurring and local distortion but without obvious blocky artifacts. The last row shows the results in the case of sampling factor 4. With increase of QP, the blur of edges grows more severe and the distortion becomes obvious, but there are still no serious blocky artifacts. The visual comparison demonstrates that CNN-US and CNN-QE effectively suppress blocky artifacts and thus the proposed method is effective in video compression using multi-sensor collaboration.

C. QUANTITATIVE MEASUREMENT

In video coding, Bjøntegaard-Delta (BD) rate [34] and rate-distortion (RD) curve [35] are usually used to evaluate the rate-distortion performance of different video encoders and BD rate can be calculated from RD curve. Both BD rate and RD curve can intuitively represent the coding efficiency improvement of the optimized algorithm compared with the original algorithm under the same video quality. A negative BD rate indicates that the coding performance of the optimized algorithm has been improved. For RD curve, higher curve points indicate better performance. We adopt two metrics to assess the proposed method. In addition, 3D-HEVC utlizes multi-view coding structure, which can make use of the information of the first view to eliminate redundancy, thus the first view is the pivotal coding content. To verify the effectiveness of the proposed method, we perform the evaluation focusing on the first view in our experiments.

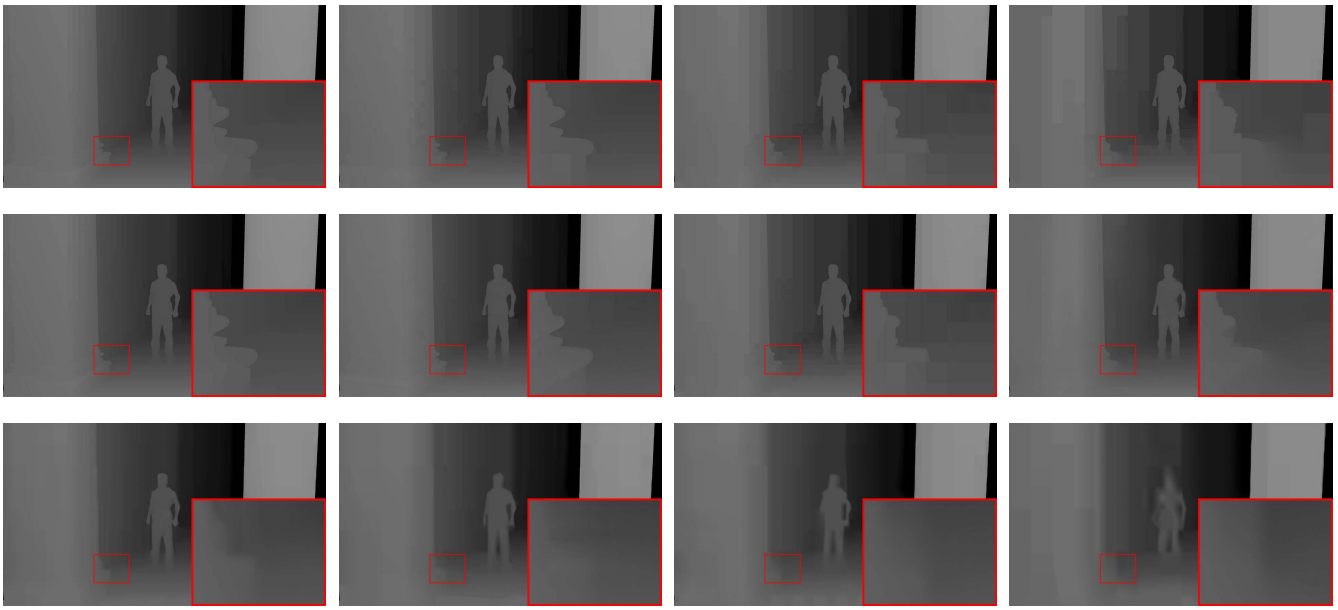

FIGURE 5. Visual comparison on *Undo_Dancer* sequence. QPs in four columns (left to right) are 34, 39, 42 and 45 in turn. The first row represents 3D-HEVC anchor. The second row and the last row represent the experimental results under sampling factors 1 and 4, respectively. The results show that the proposed method can remove blocky artifacts with local distortion in some cases.

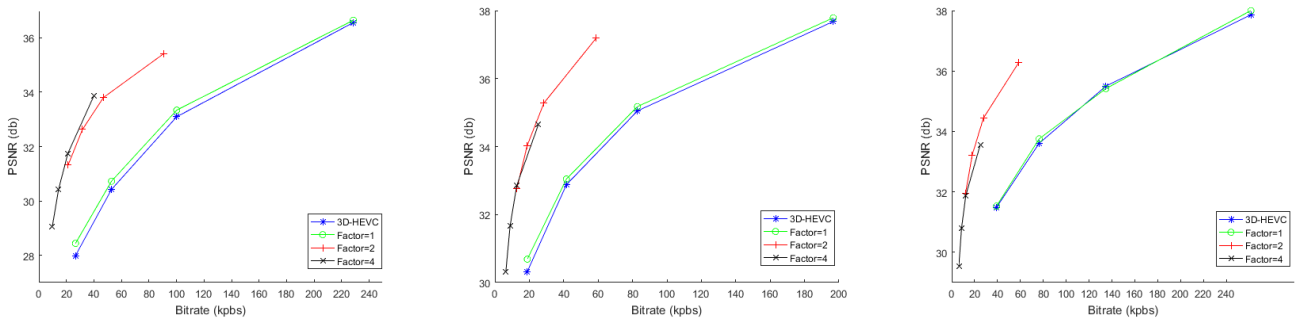

FIGURE 6. RD curves for the first view of *Kendo*, *Balloons* and *Newspaper* sequences. Left to right: *Kendo*, *Balloons* and *Newspaper*. The horizontal axis represents the bitrate for the first view of depth videos. The vertical axis represents the average PSNR value of all depth frames for the first view of depth videos. RD curves are drawn based on four QPs: 34, 39, 42 and 45. The blue line indicates 3D-HEVC anchor, and the green line, the red line and the black line represent the experimental results under sampling factors 1, 2, and 4, respectively. RD curves show that the proposed method can remarkably save bits while maintaining video quality.

Table 1 shows that BD rate results for 7 test sequences. Compared with factor 1, BD rate is significantly improved in factors 2 and 4. In factor 1, we adopted CNN-QE to achieve quality enhancement on the decoding results of 3D-HEVC and BD rate has a certain gain. In factors 2 and 4, we introduce down-sampling operation for depth videos and thus the proposed method can save bits remarkably while achieving quality enhancement. That is, the proposed method achieves a significant improvement in BD rate. Compared with factor 2, factor 4 saves more bits with a more gain in BD rate. RD curves are shown in Figs. 6 and 7, which compare the performance of the proposed method under different sampling factors in comparison with 3D-HEVC anchor. The RD curves indicate that:

1) Under sampling factor 1, CNN-QE can achieve a certain degree of quality enhancement on depth videos by the model selection strategy;

2) Under sampling factor 2, CNN-US and CNN-QE remarkably save bits while improving the quality of depth images;

3) Under sampling factor 4, CNN-US and CNN-QE remarkably save bits with a limit of quality improvement in high bitrate due to the lack of information in the input depth videos.

The visual comparison on *Kendo* and *Undo_Dancer* sequences indicates that the proposed method successfully removes blocky artifacts, and CNN-US and CNN-QE are able to perform super-resolution and quality enhancement well. The quantitative measurements on BD rate and RD curve

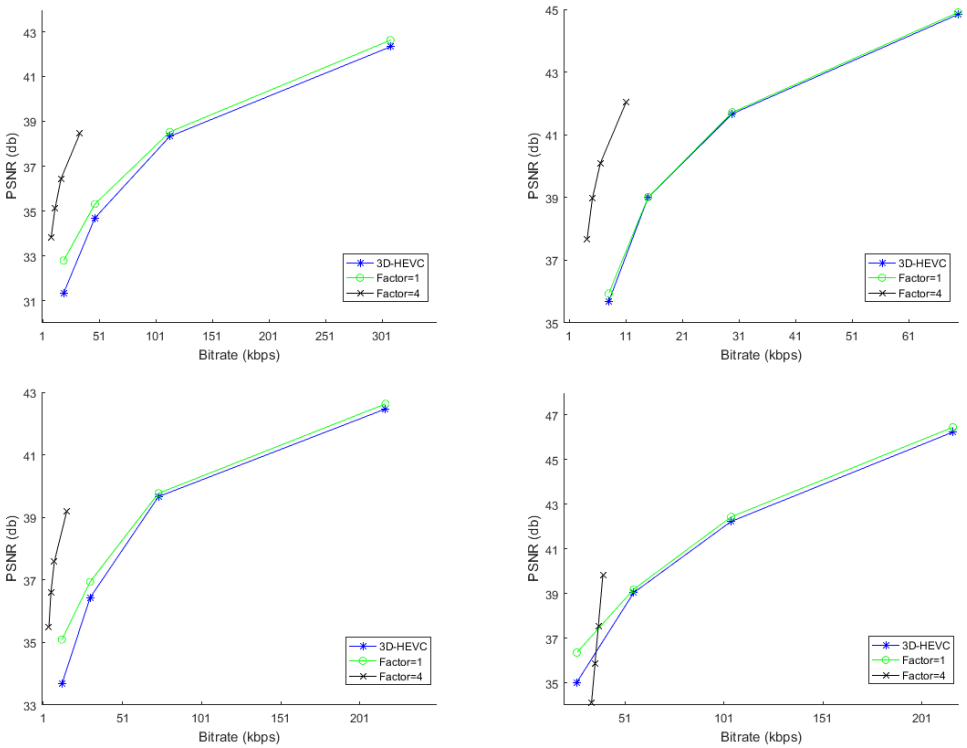

FIGURE 7. RD curves for the first view of *GT_Fly*, *Poznan_Hall2*, *Poznan_Street* and *Undo_Dancer* sequences. Top left to bottom right: *GT_Fly*, *Poznan_Hall2*, *Poznan_Street* and *Undo_Dancer*. The horizontal axis represents the bitrate for the first view of depth videos. The vertical axis represents the average PSNR value of all depth frames for the first view of depth videos. RD curves are drawn based on four QPs: 34, 39, 42 and 45. The blue line indicates 3D-HEVC anchor, and the green line, the red line and the black line represent the experimental results under sampling factors 1, 2, and 4, respectively. RD curves show that the proposed method can remarkably save bits while maintaining video quality.

verify that the multi-sensor collaboration can contribute to video compression and remarkably save bits while maintaining video quality.

V. CONCLUSION

In this paper, we propose a new video coding method that saves bits using multi-sensor collaboration. Traditional video coding methods have saved bits by removing redundancy in videos. Recently, multiple types of sensors are being deployed to many solutions, and the proposed method newly attempts to save bits using multi-sensor collaboration. We have introduced multi-sensor collaboration to the 3D video coding based on color and depth sensors. We have elaborately combined color guided depth super-resolution (CNN-US and CNN-QE) with video compression and make full use of multi-sensor collaboration to save bits without degrading image quality. Experimental results demonstrate that the proposed method achieves average 5.9%, 66.8%, and 71.0% BD-rate reductions over 3D-HEVC anchor for sampling factors 1, 2 and 4, respectively.

In our future work, we would like to extend multi-sensor collaboration to various multi-sensory data compression, e.g. visible (VIS) and infrared (IR) sensors, color and near infrared (NIR) sensors, and color and LiDAR sensors [36], [37].

REFERENCES

[1] P. Tudor, "MPEG-2 video compression," *Electron. Commun. Eng. J.*, vol. 7, no. 6, pp. 257–264, Dec. 1995.
[2] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.
[3] 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.
[4] G. J. Sullivan, J. M. Boyce, Y. Chen, J.-R. Ohm, C. A. Segall, and A. Vetro, "Standardized extensions of high efficiency video coding (HEVC)," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 6, pp. 1001–1016, Dec. 2013.
[5] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC)," *Proc. IEEE*, vol. 109, no. 9, pp. 1463–1493, Sep. 2021.
[6] 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.
[7] D. Rusanovskyy, K. Müller, and A. Vetro, *Common Test Conditions of 3DV Core Experiments*, document JCT3V-D1100, ITU-T SG, 2013.
[8] Y. Li, N. Zhu, G. Yang, Y. Zhu, and X. Ding, "Self-learning residual model for fast intra CU size decision in 3D-HEVC," *Signal Process., Image Commun.*, vol. 80, Feb. 2020, Art. no. 115660.
[9] R. Zhang, K. Jia, P. Liu, and Z. Sun, "Edge-detection based fast intra-mode selection for depth map coding in 3D-HEVC," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2019, pp. 1–4.
[10] Z. Jiang, H. Yue, Y.-K. Lai, J. Yang, Y. Hou, and C. Hou, "Deep edge map guided depth super resolution," *Signal Process., Image Commun.*, vol. 90, Jan. 2021, Art. no. 116040.
[11] Z. Huang, J. Fan, S. Cheng, S. Yi, X. Wang, and H. Li, "HMS-net: Hierarchical multi-scale sparsity-invariant network for sparse depth completion," *IEEE Trans. Image Process.*, vol. 29, pp. 3429–3441, 2020.

[12] Z. Duan and C. Jung, "Joint disparity estimation and pseudo NIR generation from cross spectral image pairs," *IEEE Access*, vol. 10, pp. 7153–7163, 2022.
[13] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6526–6534.
[14] L. H. Hughes, M. Schmitt, L. Mou, Y. Wang, and X. X. Zhu, "Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 784–788, May 2018.
[15] H. Lan, Z. Ji, C. Jung, D. Zou, and M. Li, "Multi-sensor collaboration network for video compression based on wavelet decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Aug. 25, 2022, doi: 10.1109/TCSVT.2022.3201697.
[16] J. Boyce, "MPEG immersive video coding standard," *Proc. IEEE*, vol. 109, no. 9, pp. 1521–1536, Sep. 2021.
[17] D. Ferstl, M. Ruther, and H. Bischof, "Variational depth superresolution using example-based edge representations," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 513–521.
[18] S. Mandal, A. Bhavsar, and A. K. Sao, "Depth map restoration from undersampled data," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 119–134, Jan. 2017.
[19] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.
[20] K.-H. Lo, Y. Wang, and K.-L. Hua, "Edge-preserving depth map upsampling by joint trilateral filter," *IEEE Trans. Cybern.*, vol. 13, pp. 1–14, 2017.
[21] W. Liu, X. Chen, J. Yang, and Q. Wu, "Robust color guided depth map restoration," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 315–327, Jan. 2017.
[22] X. Ye, B. Sun, Z. Wang, J. Yang, R. Xu, H. Li, and B. Li, "Depth super-resolution via deep controllable slicing network," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1809–1818.
[23] Z. Jiang, H. Yue, Y.-K. Lai, J. Yang, Y. Hou, and C. Hou, "Deep edge map guided depth super resolution," *Signal Process., Image Commun.*, vol. 90, Jan. 2021, Art. no. 116040.
[24] X. Deng and P. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2020.
[25] H. Chen, X. He, C. Ren, L. Qing, and Q. Teng, "CISRDCNN: Super-resolution of compressed images using deep convolutional neural networks," *Neurocomputing*, vol. 285, pp. 204–219, Apr. 2018.
[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.*, 2016.
[27] W. Shi, J. Caballero, F. Huszar, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[28] C. Guo, C. Li, J. Guo, R. Cong, H. Fu, and P. Han, "Hierarchical features driven residual learning for depth map super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2545–2557, May 2019.
[29] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervent.*, 2015, pp. 234–241.
[30] X. Meng, X. Deng, S. Zhu, S. Liu, C. Wang, C. Chen, and B. Zeng, "MGANet: A robust model for quality enhancement of compressed video," 2018, *arXiv:1811.09150*.
[31] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," *Proc. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 612–625.
[32] D. Scharstein, H. Hirschmuller, Y. Kitajima, G. Krathwohl, N. Nesic, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel accurate ground truth," in *Proc. German Conf. Pattern Recognit.*, 2014, pp. 31–42.
[33] (Oct. 2017). *DIML RGBD*. [Online]. Available: http://diml.yonsei.ac.kr/DIML_rgbd_dataset/
[34] M. Tang, X. Chen, J. Wen, and Y. Han, "Hadamard transform-based optimized HEVC video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 3, pp. 827–839, Mar. 2019.
[35] G. Bjøntegaard, *Calculation of Average PSNR differences Between RD-Curves*, document ITU-T VCEG-M33, VCEG 13th Meeting, 2001.
[36] H. Lan, Z. Ji, C. Jung, D. Zou, and M. Li, *New Paradigm for Video Coding: Saving Bits Using Multi-Sensor Collaboration*, document ISO/IEC JTC 1/SC 29/WG 4 m57961, Oct. 2021.
[37] Z. Ji, C. Jung, D. Zou, and M. Li, *Multi-Sensor Collaboration Meets Video Compression*, document ISO/IEC JTC 1/SC 29/WG 4 m58578, Jan. 2022.

ZHE JI received the B.S. degree in electronic engineering from Xidian University, China, in 2019, where he is currently pursuing the M.S. degree. His research interests include image processing and video compression.

HUI LAN received the B.S. degree in electronic information science and technology from Northwest University, China, in 2016. She is currently pursuing the Ph.D. degree with Xidian University, China. Her research interests include depth super-resolution and video compression.

CHEOLKON JUNG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Sungkyunkwan University, Republic of Korea, in 1995, 1997, and 2002, respectively. He was a Research Staff Member with the Samsung Advanced Institute of Technology, Samsung Electronics, Republic of Korea, from 2002 to 2007. He was also a Research Professor with the School of Information and Communication Engineering, Sungkyunkwan University, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently a Full Professor and the Director of the Xidian Media Laboratory. His research interests include image and video processing, computer vision, pattern recognition, machine learning, computational photography, video coding, virtual reality, information fusion, multimedia content analysis and management, and 3DTV.

DAN ZOU received the B.S. degree in communication engineering from Northwestern Polytechnical University, China, and the M.S. degree in instrument science and engineering from Shanghai Jiao Tong University, China, in 2017 and 2020, respectively. Since 2020, she has been a Standardization Engineer at Guangdong OPPO Mobile Telecommunications Corporation Ltd., China. Her research interests include video coding and multimedia communications.

MING LI received the B.S. degree in telecommunication engineering and the Ph.D. degree in communication and information systems from Xidian University, China, in 2005 and 2010, respectively. He was a Senior Research Staff in standardization at ZTE Corporation, China, from 2010 to 2019. Since 2019, he has been a Senior Standardization Engineer at Guangdong OPPO Mobile Telecommunications Corporation Ltd., China. His research interests include video coding and multimedia communications.