IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2020, accepted May 17, 2020, date of publication May 21, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2996258

Adaptive Multi-Modality Residual Network for Compression Distorted Multi-View Depth Video Enhancement

SIQI CHEN, QIONG LIU, (Senior Member, IEEE), AND YOU YANG, (Senior Member, IEEE)
School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China
Wuhan National Laboratory of Opto-electronics, Wuhan 430074, China

Corresponding author: You Yang (yangyou@ieee.org)

This work was supported in part by the National Natural Science Foundation of China under Grant 91848107 and Grant 61971203, in part by the National Key Research and Development Program of China under Grant 2017YFC0806202, and in part by the Wuhan Science and Technology Bureau under Grant 2019020701011422.

ABSTRACT Compression distorted multi-view video plus depth (MVD) should be enhanced at the receiver side without the original signals, especially the depth maps because they describe the positioning information in 3D space and they are important for subsequent virtual view synthesis. However, challenge arises from how to exploit the contribution from multi-modality priors from neighboring viewpoints, and how to handle the gradient vanishing when textureless depth maps are involved. In this paper, we propose a multi-modality residual network to enhance the quality of compressed multi-view depth video. Taking advantage from high correlation among different viewpoints, depth maps from adjacent views are exploited as guidance for the enhancement of depth video in target view. Color frames in target view are also involved to offer the information object contours, obtaining multi-modality guidance. The proposed network is organized a deep residual network to well eliminate distortion and restore details. Because above multi-modality guidance have different correlations with target depth video and not all information can contribute to the enhancement, an adaptive skip structure is designed to further exploit the contribution from different priors appropriately. Experimental results show that our scheme outperforms other benchmarks and achieves an average 1.935 dB and 0.0227 gains on PSNR and SSIM over all test sequences, respectively. All results on objective, subjective and 3D reconstruction suggest that our method is able to provide superiority performance in practical applications.

INDEX TERMS Compression distortion, depth map, quality enhancement, residual network.

I. INTRODUCTION

Multi-view video plus depth (MVD) is the fundamental data representation of three-dimensional (3D) and interactive visual applications, including super multi-view video, free viewpoint television and virtual reality [1], [2]. For this data representation, depth video is adopted to describe the positioning information of all visible pixels in 3D space, which is crucial for both immersive viewing experiences and virtual content synthesize in interaction for end users, especially for portable terminals in case of bandwidth limited applications. In this case, compression of MVD acts an important role in handling the big data volume of such data representation for the future success of above visual applications. Lossy compression schemes of MVD have been developed through the 3D video extension of High Efficiency Video Coding (H.265/HEVC) [3]. In these schemes, however, quality of depth videos is affected by compression distortion, subsequently destroying the positioning and structural information of the object. Therefore, quality enhancement of depth videos with compression distortion is necessary. In this case, high correlation among MVD is an important characteristic worth considering, as quality enhancement takes advantage of auxiliary information from neighboring views. Especially, in *asymmetric coding* framework, which is widely used in MVD for better coding efficiency, quantization parameter (QP) and thus quality varies among viewpoints [4], [5]. Thus, benefits can be taken from those viewpoints with higher quality in the quality enhancement of depth video with lower quality, where Fig. 1 shows an example when different benefits are taken into enhancement.

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

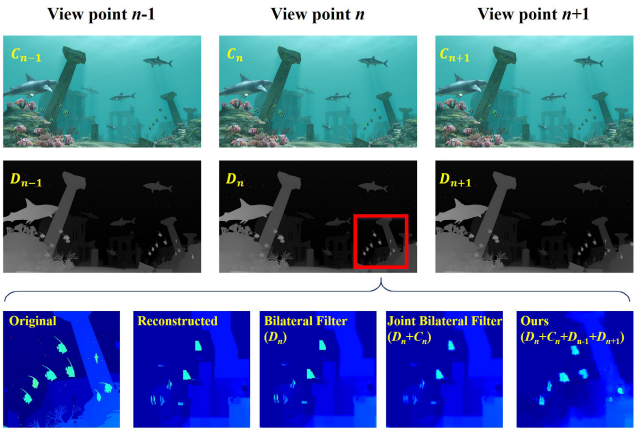


FIGURE 1. Quality enhancement on depth map D_n when priors from different viewpoints and modalities are taken.

Depth quality enhancement has witnessed a rapid development in these years, and previously proposed filters and methods have made success on this topic [6]–[9]. However, these filters are facing difficulties when compression distortions are involved in depth videos. Recently, learning-based methods have been proposed which can adaptively handle the artifacts in depth maps [10]–[12]. These works follow similar structures of networks on color image artifacts, and corresponding color images are usually taken into account as guidance. However, as for the compressed MVD, color and depth videos are both coded with different compression parameters, thus the correlation between color and depth is thus dropped. Under these concerns, more reliable auxiliary information is needed. While existing methods almost focus on auxiliary information inside mono-view, it can be observed that the depth videos of different views have high correlation and can provide reference information. Thus research on this observation is very promising for quality enhancement.

Based on the above considerations, we propose an adaptive multi-modality residual network for compressed multi-view depth video. Considering the correlation among viewpoints, depth maps from adjacent viewpoints are involved and exploited as guidance for the enhancement of depth map in current viewpoint. Since depth maps from adjacent views are with compression distortion as well, color frames in current view are also involved to offer the information of object contours. These references and guidance are combined together and regarded as multi-modality guidance. The proposed deep learning network is in the structure of residual block for better distortion elimination and details restoration. Although above multi-modality guidance has higher correlation to depth video in current view, not all information can provide positive contributions to the target of quality enhancement. Thus, an adaptive skip connection structure is designed to further exploit appropriate contributions from multi-modality guidance. In the training stage, we apply specific strategies to solve the problem caused by the characteristics of multi-view depth videos. Experimental results show that our method outperforms other state-of-the-art models, and superiority performance can be obtained on both objective and subjective evaluations.

The remainder of this paper is organized as follows. In Sec. II, we briefly survey the related works on depth video enhancement. We then present our work in details in Sec. III, and the scheme is verified in Sec. IV. Finally, we conclude our work in Sec. V.

II. RELATED WORKS

Depth maps with compression distortions can be enhanced via filters or learning based methods, and generally we category them by classical and learning-based depth map enhancement methods.

A. CLASSICAL DEPTH MAP ENHANCEMENT METHODS

In previous works, compression distortions on depth maps were processed by specific filters. The design of these filters is based on the fact that distribution of the depth value should be smooth on an object surface but sharp for boundaries. In order to recover the sharp boundaries, a candidate values based boundary filtering (CVBF) was proposed by Zhao *et al.* [6], where appropriate candidate values are selected to replace detected unreliable pixels along the boundaries. Rather than single depth map enhancement, some researchers have applied benefits of external auxiliary priors, especially color image in corresponding viewpoint are commonly used with the assumption that depth discontinuities in original depth maps are highly correlated with edges in color images. Typical example including the joint bilateral filter (JBF) [13] which originated from the bilateral filter (BF) [14], where color image is adopted as a guidance and involved into a spatial and range filter kernels. Chan *et al.* [7] further took the intrinsic noisy of real-time depth data into consideration, and an adaptive multi-lateral noise aware filtering (NAF) was proposed based on that. Different from above, Min *et al.* [15] proposed a weighted mode filtering method based on a joint histogram where color similarity between reference and neighboring pixels on the color image is used. Wang *et al.* [16] proposed an energy minimization model to emphasize the boundaries of objects in the filtering process.

B. DEEP DEPTH MAP ENHANCEMENT METHODS

Recently, convolutional neural network (CNN) showed its superior performance on low-level computer vision tasks, including the task of denoising [17], [18]. Inspired by the super-resolution CNN (SRCNN) [19], Dong *et al.* [20] designed a four layers artifacts reduction CNN (ARCNN) for effectively suppressing blocking artifacts in JPEG compressed images Based on ARCNN, a variable-filter-size residue-learning CNN (VRCNN) for artifact reduction in HEVC intra coding was proposed by Dai *et al.* [21]. Above networks were designed with shallow layers to avoid the problem of gradient vanishing. In order to tackle the problem of gradient vanishing and preserving image details in

deeper network, Mao *et al.* [22] propose a deep convolutional encoder-decoder networks with symmetric skip connections.

The residue-learning technique used in [21] can make learning process easier, more robust and converge faster. The skip connections exploited in [22] can help to back-propagate the gradients to shallow layers and pass more details to deep layers, so the network can have performance gain while the network going deeper. The success of above design on color image enhancement provide important references in the network design for depth enhancement. For the enhancement of depth frame with compression distortion, Jin *et al.* [23] cascaded the four-layer structure of ARCNN for suppressing the JPEG compression artifacts on depth frames. Considering the characteristic of depth maps, the cascaded network adopts a weighted loss function which can emphasize the edges. On the other hand, less feature can be extracted from depth maps because this kind of image is actually a less-textured gray-scale image, thus weights learned from sufficient textures are reused to initialized the learning procedure of depth maps. Considering this special characteristic of depth map, more researchers use auxiliary information in consequent methods. Li *et al.* [24] proposed the deep joint image filtering, where a CNN is utilized to construct the joint filtering approach. This framework can selectively transfer salient features that are coherent in both guidance and target images. Referring to this, Zhao *et al.* [25] designed a deep learning-based depth artifact removal method (D-ARCNN). This framework contains two sub-networks, namely joint depth-color sub-network and joint depth sub-network. In addition to color images, the gradient of color images is used in depth branch while the gradient of depth maps is stacked into color branch. Above joint methods assume that there is a co-occurrence of edges in depth map and its corresponding color image, but it is not always valid in all cases. Zhu *et al.* [26] proposed a deep residual network based on deep fusion and local linear regularization to learn the underlying correlation between depth frames and color images.

C. SUMMARY

Since complex compression parameters of MVD make enhancement more challenging, the above mono-view methods can hardly be extended to multi-view depth video quality enhancement in a straightforward way. Owing to strong representation and generalization ability, deep learning shows its superiority on our task. In this case, preserving useful details in deep network is important in the design as depth map is with less textures. In addition to network design, useful auxiliary information should be involved for better exploitation. Since MVD is assumed that the contents in different views are geometrically and semantically related, it is possible to have higher gains if cross-view depth map can be involved in the framework design. Besides, color image corresponds to the same viewpoint of depth map should be involved for sharp boundaries. However, appropriately evaluating the contribution of multi-modality priors to the quality enhancement remains a challenge in the framework design, especially when these priors are also compressed.

As noted above, the representation of MVD has unique characteristics that are different from other types of images, and a new method is necessary to satisfy these characteristics, especially when artifacts of compression distortion is considered. In our work, we dedicate to handle these problems by proposing an adaptive multi-modality residual network for compressed MVD enhancement where contributions from multi-modality priors are exploited and details can be well preserved.

III. THE PROPOSED ADAPTIVE MULTI-MODALITY RESIDUAL NETWORK

A. PROBLEM STATEMENT

Practically, depth map and color image of target (i.e., current) view is D_n and C_n at the decoder side, respectively. D_n is distorted by the process of compression, and the purpose of quality enhancement on D_n is to predict a depth map D_p^n with minimal loss to its ground truth D_{GT} as

$$\arg\min ||D_p^n - D_{GT}|| \tag{1}$$

In asymmetric coding framework, different QPs are set for higher compression efficiency, and thus comprehensive distortion levels are contained in the depth maps. This is a challenge for classical filters of depth map enhancement when distortion levels arbitrarily varied in the image. Different from that, learning-based methods are able to solve the problem of comprehensive distortion levels, but new challenge arises from how to evaluate the contributions of multi-modality priors from both depth maps of adjacent viewpoints and color images of current view. Intuitively, D_n and C_n are mutually aligned because these two images are different data representations of an exactly same scene. Therefore, scene structure is shared by these two images, and it can be only obtained through C_n instead of D_n in practical cases. For the representation of scene structure, E_n is used and it is obtained via Canny operator on C_n. For the decoder side, C_n is composed by Y, Cr and Cb component after decoded from MVD bitstream, while D_n is just represented by gray scale image which can be regarded as Y component. In this case, only Y component of C_n is picked for E_n comparing to D_n. It should be noted that textures in E_n is not the exact scene structure because one can hardly extract three-dimensional scene structure from a two-dimensional image. Therefore, both positive and negative contributions may be involved in E_n in the procedure of enhancement.

Besides the above E_n, the set of cross-view depth maps $\mathscr{D} = \{D_1, \ldots, D_{n-1}, D_{n+1}, \ldots, D_N\}$ that neighboring D_n can also contribute to the purpose of enhancement, where N is the total number of views in MVD. The reason comes form that depth maps in $\mathscr{D}$ also reveal the same scene structure of D_n but through different positions. In this case, positive contributions may come from those common regions shared by these depth maps. In order to make these common regions more clear in the enhancement, 3D warping method is applied

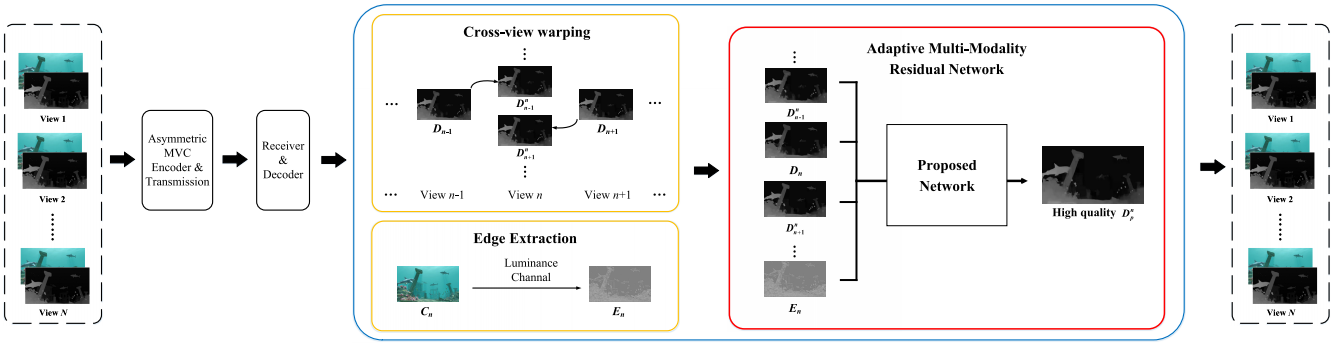


FIGURE 2. The framework of the proposed scheme for compressed multi-view depth video enhancement.

on all depth maps in $\mathscr{D}$, projecting them on the view n, to have $\mathscr{D}^n = \{D_1^n, \ldots, D_{n-1}^n, D_{n+1}^n, \ldots, D_N^n\}$.

All these contributions from both E_n and $\mathscr{D}^n$ should be properly exploited in the prediction of D_p^n in our work. Therefore, we model the prediction of D_p^n by

$$D_p^n = f(D_n, E_n, \mathscr{D}^n; W) \tag{2}$$

where $f(\cdot; W)$ is a network with weights W.

When practical application background is considered, the framework of our proposed scheme is illustrated in Fig. 2. The proposed work is a post-processing within a MVD communication system for the purpose of quality enhancement from compression distortions. At the source side, MVD is compressed by an asymmetric video encoder for better compression performance over all views. At the decoder side, the de-compressed MVD data are used as input to our work, where distortion level varies among views. Depth map D_n with lower quality is then enhanced by the proposed adaptive multi-modality residual network, and D_p^n is obtained as higher quality for output. Details of the proposed network is discussed in subsequent subsections.

B. ADAPTIVE MULTI-MODALITY RESIDUAL NETWORK

1) NETWORK STRUCTURE DESIGN

The target depth map D_n, texture map E_n and warped depth maps $\mathscr{D}^n$ are all involved as input to the proposed network for prediction of D_p^n. Deeper networks are benefit for more features and thus better performance of enhancement, but challenge arises from gradient vanishing, especially when depth map is applied as the target because this kind of images are lack of features. Therefore, residual network is adopted in our work to handle the problem of gradient vanishing and then deeper network can be used.

The proposed network is shown in Fig. 3. As depicted in this figure, residual block based network structure is applied to handle the problem of gradient vanishing. Residual block with skip connection design can address the degradation problem when deeper network is used. Different from high-level task which are insensitive to the details of images, our target is for accurate depth value and restoring the damaged information. In this case, batch normalization (BN) layers in original residual blocks is not adopted in the proposed network, because the process of feature normalization is not necessary for our task.

For more details of the residual network, each convolutional layer has 64 feature maps, while the final layer yields 1, and the size of the convolution kernel is 3×3 for all layers. We take each image as one way of input to the network, followed by a convolutional layer to extract the feature maps. The feature maps from all input will be concatenated and fed into the subsequent residual network. The network consists of 5 residual block in total. Different from original ResNet, the proposed network can hardly have a deeper design because of the high similarity and smoothness of depth maps. In the first residual block, we compute the residual between output and the target view depth branch, increasing the weight of the target view depth map. Since the target depth maps and multi-modality priors that employed as input are of different characteristics and quality, it is necessary to exploit the contributions from different modality data among E_n and $\mathscr{D}^n$. Therefore, an adaptive skip connection is applied in our network which is described in the subsequent subsection.

2) ADAPTIVE SKIP CONNECTION

Specialized in the task of depth map artifacts reduction, depth maps in target view are distorted in the process of compression, and thus high frequencies (i.e., structural information) in depth map are vanished. In this case, shallow layers are insufficient for feature extraction for subsequent enhancement. Therefore, more layers are used in our work. However, at the meanwhile, possible details or structural information may be lost in the deeper layers. In order to pass these details to deeper layers for better performance in enhancement, skip connections are used to adaptively evaluate the contribution from multi-modality priors. It should be noted that the priors from E_n and $\mathscr{D}^n$ are differently treated. For the texture map E_n, as aforementioned, both positive and negative contributions may be provided by it. For more specifically, the problem of texture copying may be found if skip connection is used for E_n, where object contours are enhanced while fake textures on object surface can also be transferred to D_p^n. Different from E_n, features are only

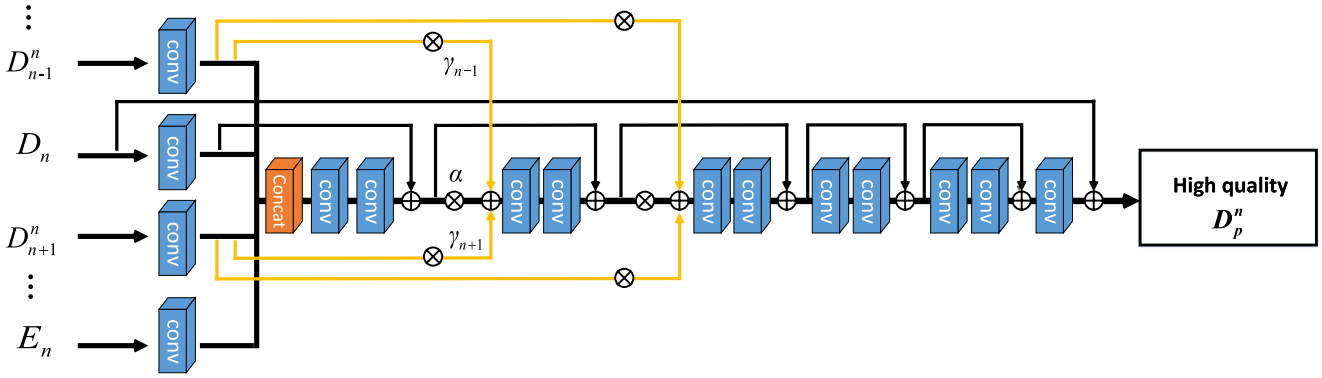


FIGURE 3. The architecture of the proposed adaptive multi-view multi-modality network for depth map enhancement from compression distortion.

extracted on common regions in all depth map of $\mathscr{D}^n$, and these features are positive contributions to D^n_p. Therefore, skip connections are necessary for these depth maps.

We summary the adaptive skip connections as following

$$y = F(\alpha x + \sum_{i=1, i\neq n}^{N} \gamma_i x_{D^n_i}, W) + x \tag{3}$$

where x and y is the input and output of residual block, respectively, and $x_{D^n_i}$ is feature obtained from D^n_i. The function $F = (\cdot, W)$ represents the residual mapping to be learned with network weights W. Parameters α and γ_i are corresponding tuning factors among x and $x_{D^n_i}$, which are learned together with other network parameters by network itself. The parameter α and γ_i is initialized as 0.5 and $0.5/(N-1)$, respectively, and they will be updated in the iteration of network solving.

Although depth features from adjacent views are beneficial for enhancement, the warping errors in $\mathscr{D}^n$ may bring artifacts on D^n_p when these features are transferred to those blocks close to the network output. Thus, the proposed adaptive skip connection is deployed on both second and third residual blocks instead of last two.

C. TRAINING STRATEGIES FOR DISTORTION MULTIPLEXED DATASET

Besides the challenge on prior exploitation of $\mathscr{D}^n$, another challenge arises from how can we efficiently train the proposed network for practical visual applications when video compression tools are applied. In the framework of asymmetric coding, QP settings are varied in the procedure of compression, resulting in different quality arrangements among views in viewpoint dimension, frames in temporal dimension, and even blocks in spatial dimension. Therefore, training strategies adopted in previous works, where JPEG compression distortions were mainly considered, can hardly be applied in our work. Different from asymmetric coding framework, QP setting in JPEG compression is uniformed, resulting in similar distortion levels among different image blocks. In this case, dataset of training can be arranged according to QP settings in JPEG oriented tasks, and at the mean time, the capability of generalization for these networks are limited due to this kind of settings. However, as discussed above, dataset of training should be arranged on multi QP settings when asymmetric coding framework is evaluated, where multiplexed distortions of different levels are simultaneously considered. Based on this dataset, the capability of generalization can be increased for the proposed network.

When multiplexed distortions are involved in the dataset, special treatments on parameter initialization and network optimization should be considered as below.

1) INITIALIZATION STRATEGY

For neural network, initialization strategy is important in the determination on whether the network can be converged, especially when the network is deep and the variation of data is significant in dataset. Specified in our work, MSRA initialization strategy proposed in [27] is applied to initial the weights and biases, which is designed to keep the input and output distribution consistent considering the characteristic of ReLU particularly. This strategy initializes the weights by drawing them from a zero-mean Gaussian distribution whose standard deviation is $\sqrt{2/n_l}$, where n_l is the number of connections of a response from l-th layer.

2) OPTIMIZATION STRATEGY

Smoothness of depth map is another challenge in the procedure of training. As aforementioned, depth map is lack of texture in most of regions, especially when higher QP is applied on it. Because of that, the gradient may be vanished soon in optimization, especially when depth map is split into small blocks in training. In order to solve this problem, Adam optimization algorithm in [28] is adopted in our work for its appropriate performance in case of sparse gradients.

IV. EXPERIMENTS AND DISCUSSIONS

A. DATASETS AND EXPERIMENT SETTINGS

1) DATASETS AND CODING PARAMETERS

There 6 in total MVD test sequences are used in our experiments with different resolutions, including Shark, Dancer, GhostTownFly, PoznanStreet, Lovebird and Balloon from the MPEG 3DV Group. These test sequences are widely used

TABLE 1. Details of depth sequences.

Sequence	View ID	Characteristics
Dancer	1, 5, 9	Software generated, simple scene, less details
Shark	1, 5, 9	Software generated, complex scene, more details
GhostTownFly	1, 5, 9	Software generated, complex scene, more details
Lovebird	4, 6, 8	Stereo matched, simple scene, less details
Balloon	1, 3, 5	Stereo matched, complex scene, more details
Street	3, 4, 5	Stereo matched, complex scene, less details

TABLE 2. Settings in asymmetric compression framework.

Parameters	Settings
Compression tools	H.265/HEVC HTM 16.0
Number of views	3
Prediction mode	Hierarchical B
Inter-view prediction	P-I-P
VSO	On
QP of the middle view	46, 47 and 48
QP of the left and right view	38, 39 and 40

in the community. Characteristics of these sequences are summarized in Table 1. Both the number of viewpoints and the viewpoint selections are suggested by the common test conditions of 3DV core experiments described in [29]. In this case, N is 3 for $\mathscr{D}$.

H.265/HEVC HTM 16.0 is used to for the depth video under the framework of asymmetric coding, and details are presented in Table 2. In each sequence, 100 consecutive frames are coded, and the middle viewpoint is set as the D_n because it is with lower quality than its neighbors.

2) DETAILS OF TRAINING AND BENCHMARKS

It should be noted that there is no available datasets can be selected for our work. In this case, the reconstructed test sequences are used as the datasets in training stage. As described above, 6 test sequences with 100 frames in each are available in the experiments. In this case, the strategy of *Leave One Out* is used in this section of experiments to avoid overfitting where 1 of 6 sequences is picked as the test while the other 5 are used in training. We present the results when all these 6 sequences are tested in the subsequent discussions. On the other hand, we train a single network with multiple QPs (i.e., multiplexed distortions), which is different from previous methods that trained a network separately for each compression level. Depth maps from 5 sequences with 3 different QPs together are used in the training set. For more specifically, we take 1 from every 10 frames in each reconstructed sequence to increase the variation of features from reconstructed depth maps. All the depth maps in training are split into blocks with size of 64×64 and with a stride of 32. Zero-padding technique is used to keep the image size during the training process. We implement our proposed network with the Caffe [30] framework and train them using NVIDIA GTX1080Ti GPUs. Training samples are randomly shuffled and the mini-batch size is 64. The weight decay is set to 0.0005 and learning rate is 10^{-5}. For Adam optimizer, we set $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$ as default.

Complehensive benchmarks are used in the comparison, including Bilateral filter (BF) [14], joint bilateral filter (JBF) [13], noise aware filtering (NAF) [7], weighted mode filtering (WMF) [15], candidate values based boundary filtering (CVBF) [6], REDNet [22]and variable-filter-size residue-learning CNN (VRCNN) [21]. Among these benchmarks, BF, JBF, NAF, WMF and CVBF are filters, while REDNet and VRCNN are learning based methods. Moreover, CVBF and VRCNN are designed specifically for HEVC scheme, and REDNet has excellent generalization ability that can deal with different low-level tasks. In the comparison, all parameters of benchmarks have been set to to have optimal performance on all the involved test sequences. For JBF and NAF, $\sigma_d = 0.1$ for depth range kernel, $\sigma_c = 0.1$ for color range kernel, and $\sigma_s = 3$ for spatial kernel are applied. For CVBF, the appropriate filtering window radium r is given in [6], and source codes released by authors are used. Since VRCNN and REDNet are not designed for HEVC compressed depth maps, we cannot use the test modal for comparison directly. In this case, VRCNN and REDNet are re-trained via the same training strategies used by our network. Because of the specific characteristics of our dataset, Gaussian initialization originally used in REDNet cannot converge anymore. Thus, we apply MSRA initialization as ours in the training procedure of REDNet.

B. OBJECTIVE EVALUATION OF DEPTH VIDEO

Two metrics are used to measure the quality of the enhanced depth maps. One is the peak signal-to-noise ratio (PSNR), which measures the pixel fidelity of depth maps. The other is the structural similarity (SSIM) [31] index, which is used to measure similarities between original and reconstructed depth maps. We use the gain between enhanced depth map and compressed depth map of PSNR and SSIM as objective evaluation result, to better show the enhancement performance of different methods.The objective results of 10-frame average PSNR and SSIM gain on each test sequence with different QPs are presented in Tables 3 and 4 respectively. As presented, the quality of enhanced depth maps obtaining by compared methods tends to be worse when QP is increasing, owing to more severe distortion of the depth maps. Instead, our method has better performance with larger QP.

Besides, not only the performance of the traditional single depth enhancement method CVBF, but also the performance of learning based method is even poorer than traditional joint method such as NAF and JBF in some test sequences. Whether shallow network like VRCNN or deep network with skip structure like REDNet, the details cannot be learned from compressed depth maps itself anymore in case of severe distortion.

In an ideal situation, a network model which is trained with abundant training samples composed of adequate characteristics can achieve good performance in all kinds of test sequences. However, in case of the lack of multi-view samples, the test performances differ in learning based methods, especially in two compared methods. In some specific test

TABLE 3. The average PSNR gain of the test sequences in comparison.

Sequence	QP	NAF	WMF	BF	JBF	CVBF	VRCNN	REDNet	ours	Avg. Gain
Shark	46	0.156	-0.113	0.181	0.533	-0.021	0.283	0.543	**1.268**	1.637
	47	0.143	-0.059	0.147	0.332	-0.015	0.261	0.408	**1.618**	
	48	0.144	-0.014	0.148	0.394	-0.009	0.264	0.455	**2.024**	
PonzanStreet	46	0.431	-0.070	0.443	0.557	-0.020	-0.304	1.096	**1.548**	1.880
	47	0.345	-0.114	0.350	0.439	-0.020	-0.489	0.729	**1.417**	
	48	0.285	-0.061	0.290	0.420	-0.011	0.226	1.907	**2.674**	
GhostTownFly	46	0.243	-0.160	0.259	0.366	-0.013	0.216	0.682	**3.084**	3.413
	47	0.187	-0.089	0.191	0.363	-0.011	0.262	0.756	**3.378**	
	48	0.208	-0.033	0.207	0.416	-0.004	0.385	0.992	**3.777**	
Dancer	46	0.287	-0.443	0.292	0.650	-0.061	-1.566	-1.598	**1.606**	2.747
	47	0.182	-0.320	0.223	0.519	-0.053	-0.677	-0.718	**2.896**	
	48	-0.001	-0.414	0.180	0.494	-0.027	-0.372	-0.410	**3.740**	
Lovebird	46	0.263	-0.077	0.263	0.358	-0.024	0.319	0.353	**1.329**	2.344
	47	0.130	-0.065	0.129	0.178	-0.011	0.187	0.466	**2.834**	
	48	0.159	-0.025	0.159	0.228	-0.012	0.211	0.566	**2.869**	
Balloon	46	0.305	-0.324	0.304	**0.504**	-0.059	0.363	-0.239	-1.391	-0.410
	47	0.239	-0.249	0.238	**0.351**	-0.049	0.299	-0.005	-0.377	
	48	0.215	-0.159	0.215	0.403	-0.038	0.366	0.368	**0.539**	
Total average gain										1.935

TABLE 4. The average SSIM gain of the test sequences in comparison.

Sequence	QP	NAF	WMF	BF	JBF	CVBF	VRCNN	REDNet	ours	Avg. Gain
Shark	46	0.0077	0.0009	0.0079	0.0096	-0.0003	0.0111	0.0254	**0.0717**	0.0572
	47	0.0086	0.0008	0.0086	0.0101	-0.0003	0.0121	0.0244	**0.0497**	
	48	0.0078	0.0013	0.0078	0.0094	-0.0002	0.0107	0.0184	**0.0501**	
PonzanStreet	46	0.0092	-0.0016	0.0100	0.0110	-0.0004	0.0085	0.0101	**0.0178**	0.0152
	47	0.0090	-0.0024	0.0094	0.0102	-0.0005	0.0068	0.0093	**0.0138**	
	48	0.0075	-0.0016	0.0082	0.0093	-0.0003	0.0055	0.0092	**0.0141**	
GhostTownFly	46	0.0155	-0.0015	0.0157	0.0173	-0.0001	0.0056	0.0159	**0.0233**	0.0328
	47	0.0163	-0.0011	0.0165	0.0187	-0.0004	0.0130	0.0224	**0.0371**	
	48	0.0167	0.0003	0.0167	0.0192	-0.0001	0.0106	0.0259	**0.0381**	
Dancer	46	0.0062	-0.0003	0.0063	**0.0072**	-0.0002	-0.0010	-0.0038	0.0044	0.0085
	47	0.0080	0.0000	0.0081	0.0095	-0.0004	0.0019	-0.0013	**0.0100**	
	48	0.0080	0.0001	0.0082	0.0096	-0.0003	0.0022	-0.0010	**0.0111**	
Lovebird	46	0.0058	0.0001	0.0058	0.0060	-0.0001	0.0047	0.0054	**0.0063**	0.0119
	47	0.0042	-0.0001	0.0042	0.0044	-0.0002	0.0036	0.0061	**0.0139**	
	48	0.0047	0.0003	0.0047	0.0051	-0.0001	0.0042	0.0072	**0.0155**	
Balloon	46	**0.0125**	-0.0030	0.0125	0.0121	-0.0012	0.0080	0.0063	0.0075	0.0104
	47	**0.0121**	-0.0034	0.0121	0.0114	-0.0012	0.0085	0.0081	0.0102	
	48	0.0116	-0.0018	0.0116	0.0117	-0.0009	0.0090	0.0093	**0.0134**	
Total average gain										0.0227

sequences, they are totally ineffective to enhance the depth maps quality. Due to the appropriate employment of multiple priors, our method works well in most test sequences. In other words, our designed network has better generalization capability compared to other learning based methods. It is worth mentioning that the reason why our method is failed to work on Balloon is the characteristics of this sequence. Different from other sequences, the depth values of a large area of the left and right views do not match the values of the target middle view. Thus, while the SSIM increases as the structure of objects is recovered, the PSNR reduces because of the wrong depth values of adjacent views. In this situation, our methods can still achieve bast performance on Balloon with QP $= 48$ which is with extremely severe distortion, for the quality gain obtained by our methods is better than the impact of mistake depth guidance. It demonstrates that our method can well handle compression distortion on high bit rates.

C. SUBJECTIVE EVALUATION OF DEPTH VIDEO

We also compare the subject results between our network and other methods, which is shown as hot maps for better visualization in Fig. 4. We utilize the first frame of test sequences with QP $= 48$ for comparison. A typical region from each depth map is marked by rectangles and enlarged to display more restored details. The depth maps illustrated in Fig. 4 are Dancer, GhostTownFly, Lovebird, Shark, PoznanStreet and Balloon from top to the bottom line.

As can be found in these results, our method can remove blocking artifacts effectively and restore more details compared with other methods. Taking the test sequence PoznanStreet and Balloon as examples, blocking artifacts appear on the car in PoznanStreet and the balloons in Balloon due to compression distortion in Fig. 4(b). Only our method reduces such blocking artifacts, restoring the flat area inside the object. More obviously, in test sequence Shark, the small fish almost completely disappear in case of severe quantization distortion, which are hardly recovered from the compressed depth map itself. Even color image is used as guidance in some filter methods, effect of color image is suppressed because of the great difference between depth map and color image. Those fish recovered exclusively by our method illustrate the contributions of the selected priors to the final performance.

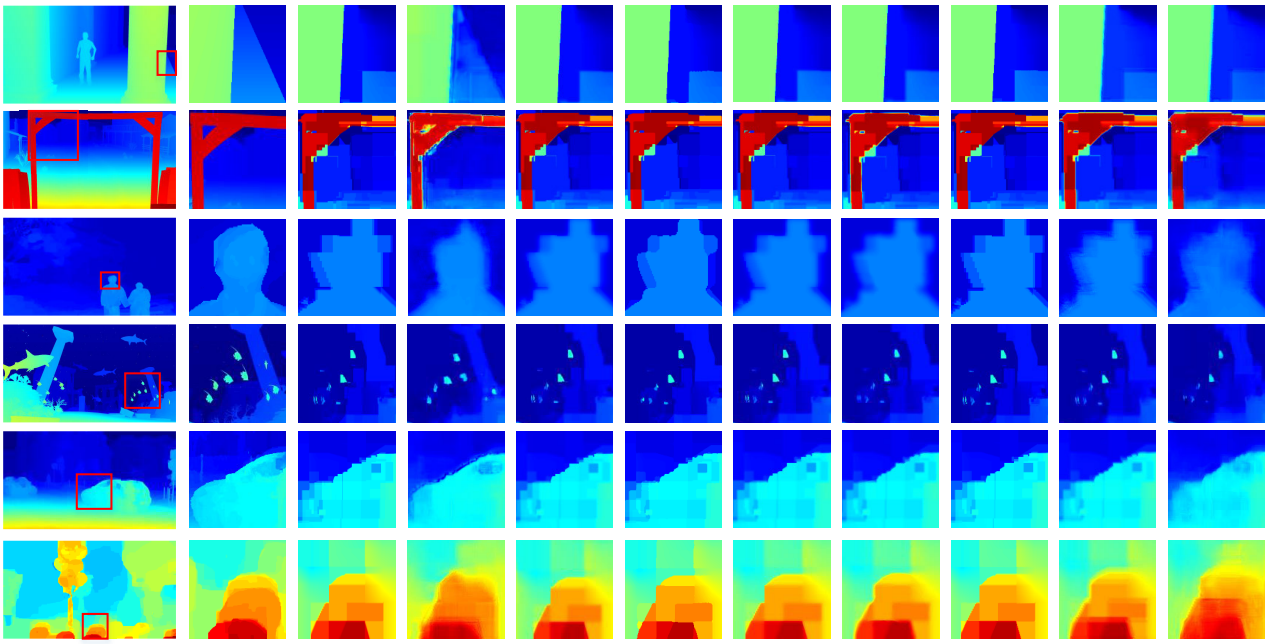


FIGURE 4. Comparisons on subjective results of the depth enhancement with different methods.

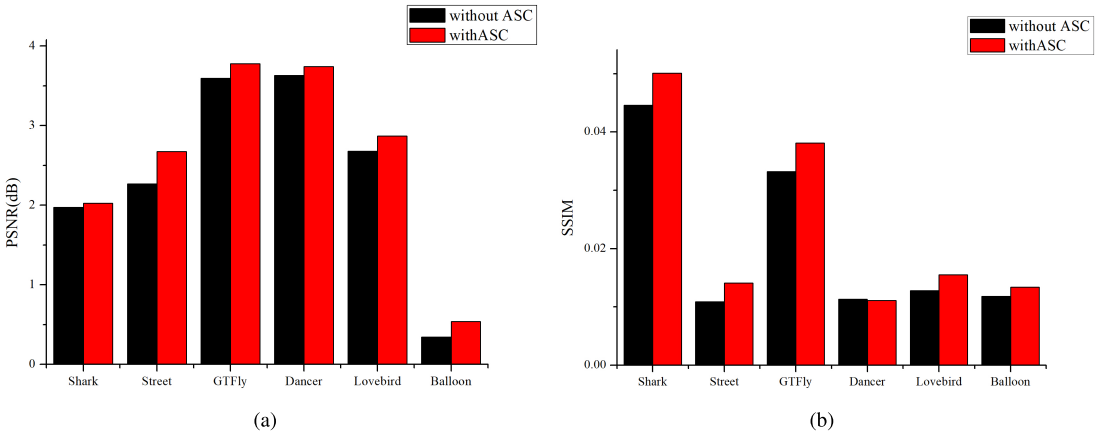


FIGURE 5. Performance evaluation between networks without/with adaptive skip connection. (a) Evaluation on PSNR gain. (b) Evaluation on SSIM gain.

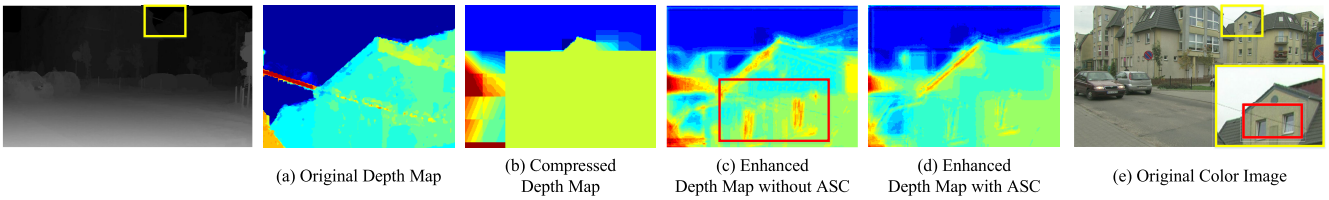
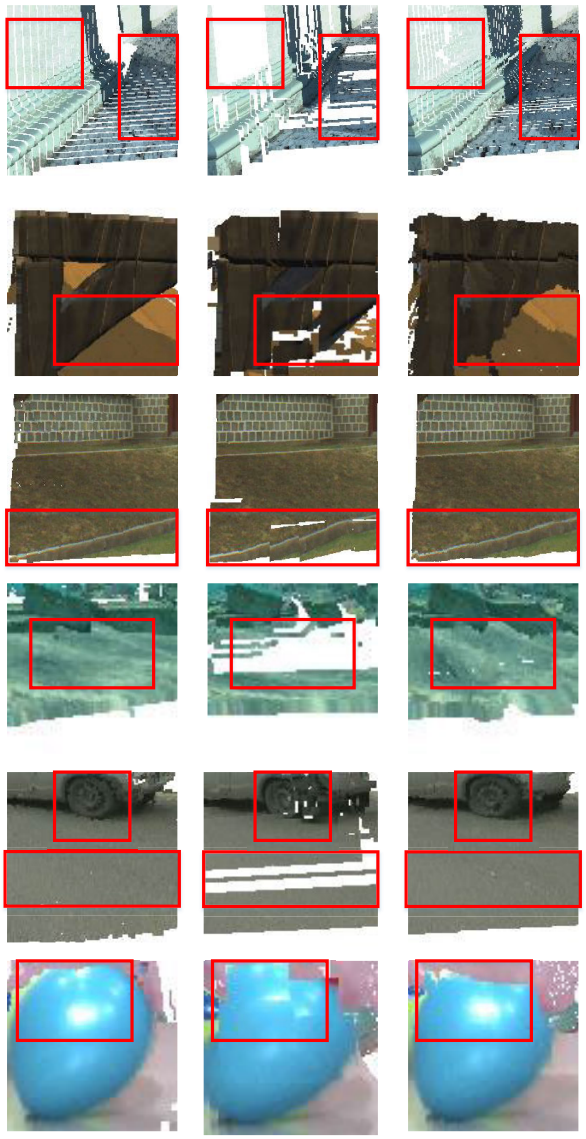


FIGURE 6. Subjective results comparison of adaptive skip connection.

D. EVALUATION OF ADAPTIVE SKIP CONNECTION

To further evaluation the performance of proposed adaptive skip connection, we trained two models that network with and without adaptive skip connection respectively. The comparison results of 10-frame average PSNR and SSIM gain on each sequence with QP48 are presented in shown in Fig. 5. As we can see, network with adaptive skip connection structure can achieve better performance on both PSNR and SSIM.



(a) Point Cloud with Original Depth Map
(b) Point Cloud with Compressed Depth Map
(c) Point Cloud with Our Enhanced Depth Map

FIGURE 7. Comparisons on perceptual quality of point cloud.

We compare the subject results for PonzanStreet as an example to more intuitively investigate where the gain comes from, and the result is shown in Fig. 6. Compared to other depth map, slight texture appear on enhanced depth map by the method without adaptive skip connection (Fig. 6(c)), which is not belong to original depth map (Fig. 6(a)). Instead, this kind of texture can be find as windows in corresponding original color image (Fig. 6(e)). In other words, texture of color image is copied to depth map. As we can see in Fig. 6(d), the addition of adaptive skip connection can suppress the texture copying problem, from which greater quality gain is obtained.

E. SUBJECTIVE EVALUATION OF POINT CLOUDS

As the quality of a depth map is fundamental to point cloud and 3D modeling, we further verify our method by this experiment. To illustrate the advantages of the enhanced depth videos, we merge original color image and corresponding depth map into a point cloud. We use Meshlab as a visualization tool to show the obtained point cloud. The results are shown in Fig. 7. The images illustrated in Fig. 7 are Dancer, GhostTownFly, Lovebird, Shark, PoznanStreet and Balloon from top to the bottom line.

As is shown, smooth object surfaces such as wall in Dancer and ground in Dancer, Shark, PoznanStreet disappear in the point clouds built by compressed depth map (Fig. 7(b)) compared to the point cloud built by original depth map (Fig. 7(a)), which results from quantization distortion. Also, object structures such as wheel in PoznanStreet and step in Lovebird are broken due to quantization distortion. With our enhanced depth map, above disappeared surfaces and broken structures are properly restored as shown in Fig. 7(c). It demonstrates that point cloud reconstruction can benefit from our proposed method, and then capabilities of high quality remote 3D applications can be improved.

V. CONCLUSION

In this paper, we propose an adaptive multi-modality residual network for depth map enhancement that distorted by compression. In this framework, depth maps from adjacent views and corresponding color images of target depth maps are taken as multi-modality priors. These priors make appropriate contribution to the enhancement due to the designed adaptive skip structure. Training set contains images with multiple QPs, which provides various compression artifacts to the network training in order to obtain a more generalized model. Experimental results show that our method outperforms the state-of-the-art methods in both the objective and subjective quality, and we also have superior performance in point cloud reconstruction.

REFERENCES

[1] D. Liu, P. An, R. Ma, W. Zhan, and L. Ai, "Scalable omnidirectional video coding for real-time virtual reality applications," *IEEE Access*, vol. 6, pp. 56323–56332, 2018.
[2] M. Tanimoto, M. Panahpour Tehrani, T. Fujii, and T. Yendo, "FTV for 3-D spatial communication," *Proc. IEEE*, vol. 100, no. 4, pp. 905–917, Apr. 2012.
[3] 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.
[4] C. Fehn, P. Kauff, S. Cho, H. Kwon, N. Hur, and J. Kim, "Asymmetric coding of stereoscopic video for transmission over T-DMB," in *Proc. 3DTV Conf.*, May 2007, pp. 1–4.
[5] S. A. Fezza and M.-C. Larabi, "Perceptually driven nonuniform asymmetric coding of stereoscopic 3D video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2231–2245, Oct. 2017.
[6] L. Zhao, A. Wang, B. Zeng, and Y. Wu, "Candidate value-based boundary filtering for compressed depth images," *Electron. Lett.*, vol. 51, no. 3, pp. 224–226, Feb. 2015.
[7] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *Proc. ECCV Workshop Multi-Camera Multi-Modal Sensor Fusion Algorithms Appl.*, 2008, pp. 1–12.
[8] W. Liu, X. Chen, J. Yang, and Q. Wu, "Variable bandwidth weighting for texture copy artifact suppression in guided depth upsampling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 10, pp. 2072–2085, Oct. 2017.

[9] Y. Yang, Q. Liu, X. He, and Z. Liu, "Cross-view multi-lateral filter for compressed multi-view depth video," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 302–315, Jan. 2019.
[10] G. Riegler, R. Ranftl, M. Rüther, T. Pock, and H. Bischof, "Depth restoration via joint training of a global regression model and CNNs," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 1–58.
[11] J. Jeon and S. Lee, "Reconstruction-based pairwise depth dataset for depth image enhancement using CNN," in *Proc. ECCV*, Sep. 2018, pp. 422–438.
[12] X. Wang, P. Zhang, Y. Zhang, L. Ma, S. Kwong, and J. Jiang, "Deep intensity guidance based compression artifacts reduction for depth map," *J. Vis. Commun. Image Represent.*, vol. 57, pp. 234–242, Nov. 2018.
[13] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," *ACM Trans. Graph.*, vol. 26, no. 3, p. 96, Jul. 2007.
[14] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. 6th Int. Conf. Comput. Vis.*, Jan. 1998, pp. 839–846.
[15] D. Min, J. Lu, and M. N. Do, "Depth video enhancement based on weighted mode filtering," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1176–1190, Mar. 2012.
[16] Y. Wang, Y. Yang, and Q. Liu, "Feature-aware trilateral filter with energy minimization for 3D mesh denoising," *IEEE Access*, vol. 8, pp. 52232–52244, 2020.
[17] X. Zhang and R. Wu, "Fast depth image denoising and enhancement using a deep convolutional network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2499–2503.
[18] W. Dong, P. Wang, W. Yin, G. Shi, F. Wu, and X. Lu, "Denoising prior driven deep neural network for image restoration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2305–2318, Oct. 2019.
[19] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.
[20] C. Dong, Y. Deng, C. C. Loy, and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 576–584.
[21] Y. Dai, D. Liu, and F. Wu, "A convolutional neural network approach for post-processing in HEVC intra coding," in *Proc. Int. Conf. Multimedia Modeling (MMM)*, in Lecture Notes in Computer Science, vol. 10132. Springer, 2017, pp. 28–39.
[22] X.-J. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. NIPS*, 2016, pp. 2802–2810.
[23] Z. Jin, L. Luo, Y. Tang, W. Zou, and X. Li, "A CNN cascade for quality enhancement of compressed depth images," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.
[24] Y. Li, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep joint image filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 154–169.
[25] L. Zhao, J. Liang, H. Bai, A. Wang, and Y. Zhao, "Convolutional neural network-based depth image artifact removal," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2438–2442.
[26] J. Zhu, J. Zhang, Y. Cao, and Z. Wang, "Image guided depth enhancement via deep fusion and local linear regularizaron," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4068–4072.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[29] K. Müller and A. Vetro, *Common Test Conditions of 3DV Core Experiments*, document JCT3V-G1100, San Jose, CA, USA, Jan. 2014.
[30] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[31] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

SIQI CHEN received the B.S. degree in information engineering from the School of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan, China, in 2018, where she is currently pursuing the master's degree. Her research interests include deep learning, 3D vision, and image processing.

QIONG LIU (Senior Member, IEEE) received the Ph.D. degree in computer science from the School of Computer Science, Wuhan University, Wuhan, China, in 2008.

From 2010 to 2012, she has worked as a Postdoctoral Fellow with the Automation Department, Tsinghua University. She is currently an Associate Professor with the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan. She is also with the Wuhan National Laboratory of Opto-Electronics, Division of Intelligent Media and Fiber Communications, Wuhan. She has authored or coauthored over 30 technical articles. She holds authorized 20 patents. Her research interests include cross-discipline researches between artificial intelligence and three-dimensional (3D) computer vision, including 3D visual communication, human–robot interaction, human intention, and interactive visual applications. She has been a Committee or TPC Member of over ten international conferences. She has also been a Reviewer of over ten prestigious international journals from the IEEE, ACM, OSA, and other associations. She was an Associate Editor of *KSII Transactions on Internet and Information Systems*, in 2016, and *IET Image Processing*, in 2019. She was a Guest Editor of *Multimedia Tools and Applications*, in 2015.

YOU YANG (Senior Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2009.

From 2009 to 2011, he has worked as a Postdoctoral Fellow with the Automation Department, Tsinghua University. From 2011 to 2013, he was a Senior Research Scientist with Sumavision Research. He is currently the Head of the Department of Information Engineering, Huazhong University of Science and Technology, Wuhan, China. He is also with the Wuhan National Laboratory of Opto-Electronics, Division of Intelligent Media and Fiber Communications, Wuhan. He has authored or coauthored over 70 peer-reviewed articles. He holds authorized 22 patents. His research interests include three-dimensional (3D) vision system and its applications, including multi-view imaging systems, 3D/VR/AR content processing and visual communications, human–machine interaction techniques, and interactive visual applications. He was elected as a Fellow of the Institute of Engineering and Technology (FIET), in 2018. He has been a Committee/TPC Member or the Session Chair of over 30 international conferences, including ICME, ICASSP, VCIP, ICIMCS, MMM, and others. He was invited to be the Judge of the IET Innovation Award, in 2019. He has also been a Reviewer of 32 prestigious international journals from IEEE, ACM, OSA, and other associations. From 2014 to 2016, he was a Guest Editor of *Neurocomputing*. He has been an Associate Editor of IEEE Access, since 2018, *IET Image Processing*, since 2018, and *PLoS ONE*, since 2015.