CONVOLUTIONAL NEURAL NETWORK-BASED DEPTH IMAGE ARTIFACT REMOVAL

Lijun Zhao [⋆] [†], *Jie Liang* [†], *Huihui Bai* [⋆], *Anhong Wang* [#], *Yao Zhao* [⋆]

[⋆]Institute Information Science, Beijing Jiaotong University,
Beijing 100044, P. R. China
[†] School of Engineering Science, Simon Fraser University,
ASB 9843, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
[#]Institute of Digital Media & Communication, Taiyuan University of Science and Technology,
Taiyuan 030024, China

ABSTRACT

In 3D video coding and depth-based image rendering, the distortion of the compressed depth image often leads to wrong 3D warpping. In this paper, by generalizing the recent work of convolutional neural network (CNN)-based depth image up-sampling, we propose a CNN-based depth image artifact removal scheme, where both the compressed depth and color images are used to enhance the depth accuracy. The proposed CNN has two sub-networks: joint depth-color sub-network and joint depth sub-network. During the depth and color feature extraction, the gradient of the depth image is used as the input to color image, while the gradient of color image is used as the input of depth feature extraction. Such an exchange of gradient information improves the learned features. Experimental results in terms of both objective and subjective quality of the depth and color images verify the efficiency of the proposed method.

Index Terms— 3D video coding, depth filtering, joint filtering, CNN

1. INTRODUCTION

Recently 3D video has become increasingly popular. It is usually represented by multiple texture views and depth maps, which is called multiview video plus depth (MVD) format [1]. The depth maps can be used to synthesize new videos of virtual viewpoints and achieve the free-viewpoint video effect by the depth-image-based rendering method (DIBR) [2, 3]. However, due to the large amount of data involved, the depth maps are usually compressed by standard video codecs such as HEVC [4] or 3D-HEVC [5]. This inveitably introduces structure losses in the reconstructed depth image, and causes deformed objects in the synthesized virtual color images. Therefore, it is necessary to improve the quality of the depth image.

Corresponding author: Huihui Bai (email: hhbai@bjtu.edu.cn). This work was supported in part by National Natural Science Foundation of China (No. 61672087, 61672373), and CCF-Tencent Open Fund.

There are two types of depth image filtering methods: single depth image filtering (SDIF) and joint depth image filtering. In [6, 7, 8], some works about SDIF have been done. For example, in [6], a depth boundary reconstruction filter is proposed, by using occurrence frequency, similarity, and closeness of pixels, but the complexity of the method is quite high. In [7], a low-complexity adaptive block truncation filter was developed to restore the sharp boundaries with adaptive block repositioning and other methods that can improve the refinement accuracy of depth values. In [8], in order to keep the sharp edges, an appropriate candidate value was chosen to replace each unreliable pixel based on the nearest reliable pixels and the mean values of neighboring region.

Compared to SDIF, joint depth image filtering uses color image information to improve the depth quality, which is similar to joint depth image super-resolution [9, 10, 11, 12]. In [13], both color frame and depth frame was utilized to suppress coding artifacts and preserve edges when large artifacts were caused by video coding. In [14], the depth-merged color image was used to refine the quality of the distorted depth image using joint iterative guidance filtering, where the smoothed color image was taken as the guidance for the quality's improvement of depth image.

The methods above use handcrafted features to improve the depth map quality. Recently, the convolutional neural network (CNN) has been proven to be a powerful tool to resolve many low-level and high-level vision problems. One typical application in low-level vision problems is CNN-based image super-resolution (SRCNN) [15], where various features were learned by CNN, and the image super-resolution was achieved by directly learning an end-to-end mapping between the low-resolution image and high-resolution image. In [16], a new architecture of CNN is used to extract the features and transfer image structure from the guidance image to the target image, which can be used for joint depth up-sampling and structure-texture separation etc. In this architecture, three SRCNN sub-networks are proposed to form one framework, namely target image feature extraction, guidance image fea-

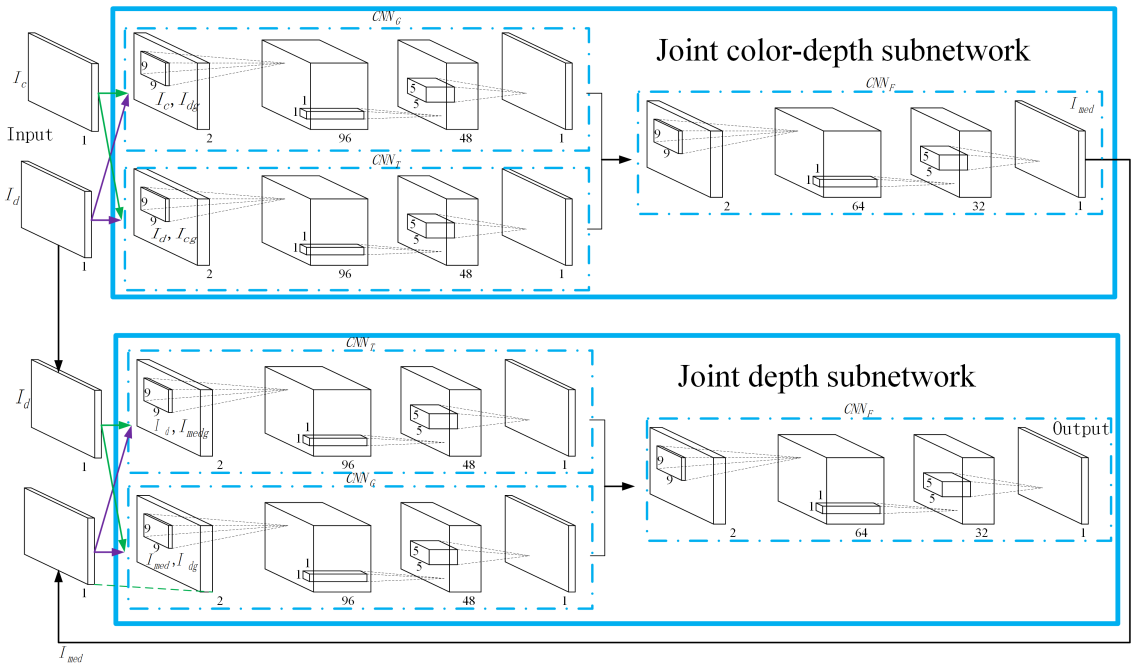

Fig. 1. The network architecture of the proposed deep learning-based depth artifact removal (D-ARCNN).

ture extraction, and final feature concentration. This network design comes from the FlowNet in [17]. Another closely related method is the artifact reduction CNN of [18] (ARCNN), which consists of four convolutional layers. For compressed depth image, not only depth artifact is required to be reduced, some structure information can also be restored by the similarity between color and depth images when the corresponding color image is given.

Following the work in [15, 16], in this paper, a CNN-based joint depth image artifact removal (D-ARCNN) framework is proposed, which includes two feature extraction stages: joint color-depth feature extraction, and joint depth feature extraction. The separate extraction strategy in [16, 17] is employed. During the separate depth and color feature extraction, the gradient of depth image is used as an input to the color image, and the gradient of color image is also added into the input of depth feature extraction. The structure of color image is aimed to be transferred into depth image to improve the accuracy of depth image.

The rest of this paper is organized as follows. In Section 2, the proposed method are presented. Section 3 presents experimental results, and conclusions and future work are discussed in Section 4.

2. PROPOSED METHOD

The proposed CNN framework is shown in Fig. 1, which includes two sub-networks: joint color-depth sub-network and joint depth sub-network. $I_c(x, y)$ is the pixel value of color image I_c at location (x, y), I_d is compressed depth image, I_{GT} is the ground truth depth image, and I_{cg} and I_{dg} are the gradients of the color image and depth image respectively. The two sub-networks have the same structure as in [16] (denoted as core network). Each sub-network has three components as in [15], namely CNN_T, CNN_G, and CNN_F. In CNN_T, the target image is taken as the network's input to extract features. The CNN_G is used to extract the features from the guidance image to enhance the target image. Finally the features from the target image and the guidance image are stacked together as the input of CNN_F to achieve joint filtering. Each of CNN_T, CNN_G, and CNN_F has three functionalities: patch extraction and representation, non-linear mapping, and reconstruction [15].

Depth images are characterized by homogeneous regions divided by clean and sharp boundaries. This property should be considered when designing neural network for them. In the proposed network, firstly the gradients of the depth image and color image are exchanged, as shown by the purple and green arrows in Fig. 1, which will benefit each other and facilitate future feature fusion. Compared with the original high-resolution image, both the low-resolution depth image and compressed depth image have the structure loss for depth super-resolution and compressed depth image filtering, but there are some differences between them, such as the type of noise and the degree of deformation in the depth image. Therefore, during the filtering of compressed depth image these differences should be considered. For example, severe structure deformation often appears except for coding artifacts when the depth image is compressed with large quantization step. So we introduce two sub-networks to filter

compressed depth image. The output of the joint color-depth sub-network is used as the initial representation I_{med} of the depth image, which is sent to the joint depth sub-network as an input. The latter combines the initial estimation I_{med}, its gradient I_{medg}, compressed depth image I_d, and its gradient I_{dg} to produce the final output. Although the two sub-networks share the same network, their inputs are slightly different, which plays a key role in improving the final depth accuracy, which will be demonstrated in the experimental results. Meanwhile, CNN_T, CNN_G, and CNN_F have the similar network except for the feature map's number difference of each layer's filtering between them. In CNN_T and CNN_G, three convolutional layers are included: first convolutional layer with size $9 \times 9 \times 2 \times 96$, second convolutional layer with $1 \times 1 \times 96 \times 48$ and last convolutional layer with $5 \times 5 \times 48 \times 1$. The first two convolutional layers of CNN_T, CNN_G, and CNN_F are respectively followed by Rectified linear unit to activate the neural.

Given training image patches $\{I_c^k, I_{cg}^k, I_d^k, I_{dg}^k\}_{k=1}^L$, the parameters of the proposed CNN are learned by minimizing the following function:

$$||I_{GT} - f(I_c, I_{cg}, I_d, I_{dg})||_2^2, \quad (1)$$

where $f(\cdot)$ is the filtering operator, and k is the index of patches with the number of L. Finally the loss function is back-propagated into all layers by stochastic gradient descent with momentum for learning the parameters. Actually, the whole network of D-ARCNN can be trained in an end-to-end way to learn the CNN parameters, but in our simulation the joint depth sub-network is simply trained after the training of the joint color-depth sub-network is finished. The initial estimation with joint color-depth sub-network is got by minimizing $||I_{GT} - f(I_c, I_{cg}, I_d, I_{dg})||_2^2$ denoted as D-ARCNN0. For the joint depth sub-network training, the parameters are learned by minimizing the following function:

$$||I_{GT} - f(I_{med}, I_{medg}, I_d, I_{dg})||_2^2. \quad (2)$$

3. EXPERIMENTAL RESULTS

3.1. The training configurations

To evaluate the efficiency of the proposed method, the intra mode of HEVC v9.0 [4] is used to compress the depth video with QP=36 and 41. From MPI Sintel color-depth dataset [19], 112000 patches are collected to train our model, in which 93300 patches are the training data and the others are used for validation. Notice that all the training and testing RGB-D data are generally normalized to be in the range of 0 to 1. The MatConvNet toolbox [20] is used to train and learning our model. The convergent curve of our D-ARCNN network for QP=41 is presented in Fig. 2 with 350 epoches. We train joint depth-color sub-network with the first 250 epoches, while joint depth sub-network is trained with last 100 epoches.

Table 1. The objective measurement of filtered depth images with different methods

	M/Seq	B	C	U	S
QP41	Coded	37.87	40.14	44.63	38.46
	JTF[13]	38.01	40.24	44.95	38.59
	DBRF[6]	35.42	38.28	43.39	35.34
	ADTF[7]	37.38	40.07	44.63	38.25
	CVBF[8]	37.42	40.06	44.70	38.10
	D-ARCNN0	37.99	38.95	44.85	39.08
	D-ARCNN	**38.07**	**40.44**	**45.21**	**39.16**
	M/Seq	B	C	U	S
QP36	Coded	40.77	44.43	48.35	42.05
	JTF[13]	40.96	44.60	48.80	42.27
	DBRF[6]	37.41	40.69	46.39	37.08
	ADTF[7]	40.28	44.15	48.14	41.74
	CVBF[8]	40.13	44.28	48.30	41.37
	D-ARCNN0	41.04	43.99	48.38	42.31
	D-ARCNN	**41.05**	**44.75**	**48.90**	**42.49**

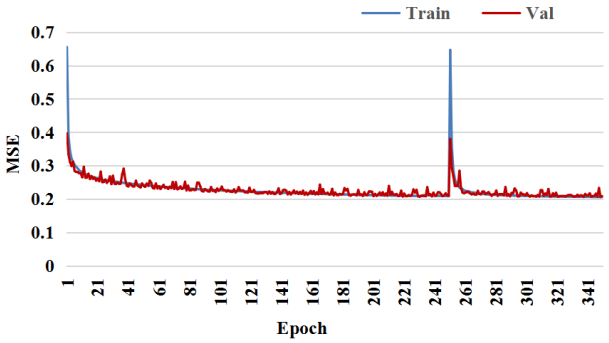

Fig. 2. The convergence of our D-ARCNN network

3.2. The comparison of depth image quality and synthesized virtual image quality

The first 100 frames of four standard 3D testing sequences in multi-view videos plus depth format are tested, including 1024×768 "Book_Arrival (Denoted as B)", 1280×960 "Champagne_Tower (C)", 1920×1088 "Undo_Dancer (U)", and "Shark (S)". Among them, "Book_Arrival " and "Champagne_Tower" are real scene sequences while the others are non-real scenes.

The depth image provides the geometry information to generate new images. To render the new viewpoint virtual images, the standard synthesis software with 1D-fast mode of 3D-HEVC [5] (HTM-DEV-2.0-dev3 version) is taken as the experimental platform when the uncompressed texture image and their compressed depth images (filtered or non-filtered) are given. Besides, before D-ARCNN filtered depth images are used for DIBR, the sharp operator, such as CVBF is advised to generate clear boundaries along the edges.

The peak signal-to-noise ratio (PSNR) is used to measure the quality of images filtered by different methods in TA-

Table 2. The objective measurement of synthesized virtual images

	M/Seq	B	C	U	S
QP41	Coded	50.17	44.09	49.56	45.4
	JTF[13]	50.35	44.04	50.2	45.63
	DBRF[6]	50.15	44.25	50.92	45.37
	ADTF[7]	50.86	44.47	51.08	46.9
	CVBF[8]	51.02	44.57	51.24	47.11
	D-ARCNN0	50.52	44.00	50.75	47.19
	D-ARCNN	**51.08**	**44.58**	**51.46**	**48.03**
	M/Seq	B	C	U	S
QP36	Coded	52.35	46.42	52.06	48.01
	JTF[13]	52.6	46.31	52.9	48.37
	DBRF[6]	51.78	45.67	53.42	47.07
	ADTF[7]	53.09	46.54	53.73	49.33
	CVBF[8]	52.99	46.43	53.97	49.55
	D-ARCNN0	52.86	46.01	53.13	49.13
	D-ARCNN	**53.14**	**46.57**	**54.31**	**50.16**

BLE 1-2, where M/Seq denotes to Method/Sequence. For depth image, the average PSNR of all the testing frames is used to measure whether the depth's quality is enhanced and how much they changes. Meanwhile, the average PSNR of color image's Y-U-V components evaluates how much depth image's changes affect the synthesized virtual image. The proposed D-ARCNN has better performance than four other methods: JTF [13], DBRF [6], ADTF [7],CVBF [8], in terms of both the depth image and sythesized virtual image's objective measurement. The visual quality comparison of both depth and color image is presented in Fig. 3, which shows that the proposed method has better performance than the other four methods.

4. CONCLUSION

In this paper, the convolution neural networks is proposed to improve the accuracy of compressed depth image by transferring shared structure in the color image. The proposed CNN includes two sub-networks: joint color-depth sub-network and joint depth sub-network. The joint depth sub-network is proposed to further enhance the depth quality after joint color-depth sub-network. During the depth and color feature extraction, the gradient of the depth image is stacked into the input of color image's feature extraction while the gradient of color image is used as the input of depth feature extraction.

The proposed CNN framework can be further improved. For example, the architecture of the joint color-depth sub-network is unbalanced, with deeper CNN for the depth and color feature extraction, or deeper network for depth and color feature concentration. Moreover, the design of loss function can be improved. A possible approach is to include the gradient information into the output loss.

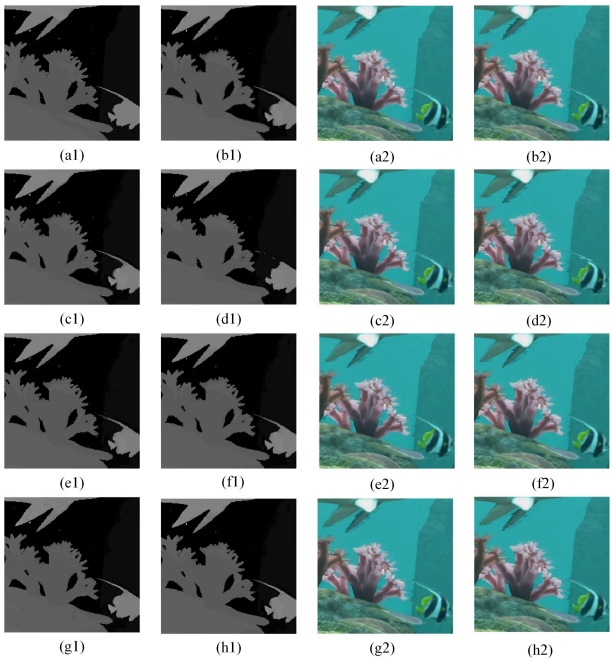

Fig. 3. (a1) Part of the original first frame of depth map for Shark in View 1; (b1) HEVC compressed depth map of (a1) with QP=41; (c1-h1) are filtered images from (b1) by JTF [13], DBRF [6], ADTF [7],CVBF [8], D-ARCNN0, and D-ARCNN, (a2)-(h2) Parts of synthesized image with one neighbouring view's depths from (a1)-(h1)

5. REFERENCES

[1] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *IEEE International Conference on Image Processing*, 2007, pp. 201–204.

[2] Fehn C., "Depth-image-based rendering, compression, and transmission for a new approach on 3d-tv," *Proc. SPIE 5291, Stereoscopic Displays and Virtual Reality Systems XI*, vol. 93, 2004.

[3] Tian D., Lai P., Gomila C., and Lopez P., "View synthesis techniques for 3d video," *Proc. SPIE 7443, Applications of Digital Image Processing XXXII, 74430T*, vol. 93, 2009.

[4] JCT-VC, "Hevc test software (hm) [online]," https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-9.0/.

[5] JCT-3V, "3d-hevc test software (htm) [online]," http://hevc.kw.bbc.co.uk/git/w/jctvc-3de.git/shortlog/refs/heads/HTM-DEV-2.0-dev3-Zhejiang.

[6] Oh K ., Vetro A., and Ho Y., "Depth coding using a boundary reconstruction filter for 3-d video systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 350–359, 2011.

[7] X. Xu, Po L., Cheung C., Cheung K., and Feng L., "Adaptive depth truncation filter for mvc based compressed depth image," *Signal Processing: Image Communication*, vol. 29, no. 3, pp. 316–331, 2014.

[8] Zhao L., Wang A., Zeng B., and Wu Y, "Candidate value-based boundary filtering for compressed depth images," *Electronics Letters*, vol. 51, no. 3, pp. 224–226, 2015.

[9] Kopf J., Cohen M., Lischinski D., and Uyttendaele M., "Joint bilateral upsampling," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 96, 2007.

[10] Ham B., M Cho, and J Ponce, "Robust image filtering using joint static and dynamic guidance," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[11] Liu W., Chen X., Yang J., and Wu Q., "Variable bandwidth weighting for texture copy artifacts suppression in guided depth upsampling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, no. 1, 2016.

[12] Jung C., Joo S.and, and Su M., "Depth map up-sampling with image decomposition," *Electronics Letters*, vol. 51, no. 22, pp. 1782–1784, 2015.

[13] Liu S., P. Lai, Tian D., and Chen C., "Joint trilateral filtering for depth map compression," *Visual Communications and Image Processing*, 2010.

[14] Zhao L., Bai H., Wang A., and Zhao Y., "Joint iterative guidance filtering for compressed depth images," in *Visual Communications and Image Processing*, 2016.

[15] Dong C., Loy C., He K., and Tang X., "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

[16] Li Y., Huang J., Ahuja N., and Yang M., "Deep joint image filtering," in *European Conference on Computer Vision*, 2016.

[17] Dosovitskiy A., Fischer P., Ilg E., Hausser P., Hazirbas C., Golkov V., and Brox T., "Flownet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 2758–276.

[18] Dong C., Deng Y., Change C., and Tang X., "Compression artifacts reduction by a deep convolutional network," in *IEEE International Conference on Computer Vision*, 2015, pp. 576–584.

[19] Butler D., Wulff J., Stanley G., and M. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conference on Computer Vision*, 2012.

[20] Vedaldi A. and Lenc K., "Matconvnet: Convolutional neural networks for matlab," in *ACM International Conference on Multimedia*, 2015, pp. 689–692.