LETTER *Special Section on Picture Coding and Image Media Processing*

Neural Network-Based Post-Processing Filter on V-PCC Attribute Frames

Keiichiro TAKADA[†a)], **Yasuaki TOKUMO**[†], **Tomohiro IKAI**[†], ***Nonmembers***, ***and*** **Takeshi CHUJOH**[†], ***Senior Member***

SUMMARY Video-based point cloud compression (V-PCC) utilizes video compression technology to efficiently encode dense point clouds providing state-of-the-art compression performance with a relatively small computation burden. V-PCC converts 3-dimensional point cloud data into three types of 2-dimensional frames, i.e., occupancy, geometry, and attribute frames, and encodes them via video compression. On the other hand, the quality of these frames may be degraded due to video compression. This paper proposes an adaptive neural network-based post-processing filter on attribute frames to alleviate the degradation problem. Furthermore, a novel training method using occupancy frames is studied. The experimental results show average BD-rate gains of 3.0%, 29.3% and 22.2% for Y, U and V respectively.

key words: *3D video coding, point cloud compression, neural network coding*

1. Introduction

Video-based point cloud compression (V-PCC) has been standardized as ISO/IEC 23090-5 by the Moving Pictures Experts Group (MPEG) and the first version was finalized [1]. V-PCC efficiently compress 3-dimensional (3D) point cloud using existing 2D video compressions, such as HEVC [2] or VVC [3].

Figure 1 shows an overview of V-PCC data. V-PCC divides a point cloud into a number of connected regions called 3D patches. Each point in the regions is projected to a plane to derive 2D patches. Derived 2D patches are collected and packed into a frame in which the following patch information and 2D image frames (video) are generated.

- **Atlas:** patch information, e.g., patch position, patch size, patch orientation, etc.
- **Geometry:** image frames representing normalized distance value between the plane and the point.
- **Occupancy:** image frames representing whether a projected point exists or not.
- **Attribute:** image frames representing projected attribute value, e.g., color or transparency, etc.

However the quality of geometry and attribute frames may be degraded due to video compression and the corresponding artifacts, block noise and banding noise in attribute images, are still visible when viewing images are rendered. Recently, neural network-based processing has been proven to be effective in achieving 8K quality video transmission with its resolution enhancement [4] and in reducing V-PCC geometry artifact with its 3D patch filtering [5]. Adaptive neural network-based methods utilizing neural network model parameter (neural network model) transmission have also been reported in 2D video compression [6], [7] where the MPEG Neural Network Coding (NNC) standard, ISO/IEC 15938-17, is used for model compression [8]. Transmitting neural network model provides adaptability for various contents and specific model is beneficial to make model complexity relatively small. Therefore, it can be effectively applied to existing standards as an out-of-loop post-processing and it is known that neural network is superior to traditional image processing [9]. This paper proposes an adaptive neural-network-based post-processing filter (post-filter) for V-PCC. The experimental result shows the coding efficiency improvement and our training method's effectiveness.

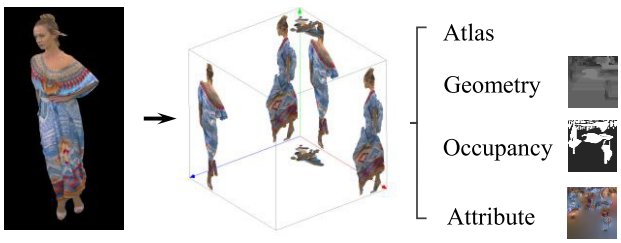

Fig. 1 Overview of V-PCC data: 3D point clouds are represented by atlas, geometry, occupancy, and attribute.

2. Proposed Method

2.1 Neural Network Post-Filter on Attribute Frames

As shown in Fig. 2, the proposed method applies a neural network-based post-filter on attribute frames as an additional process between attribute video decoding and projection back to 3D space. It is noted that attribute data is dominant in quantity in V-PCC bitstreams. For example, the average ratio of attribute bitrate to total bitrate is 52.8% on V-PCC common test conditions (CTC) [10] under the random access (RA) configuration. Therefore, enhancing attribute frames quality should effectively improve the coding efficiency of total V-PCC. This proposal transmits neural-network model using NNC. In the proposal, the model size

Manuscript received January 6, 2023.
Manuscript revised April 19, 2023.
Manuscript publicized July 13, 2023.
[†]The authors are with Sharp Corporation, Chiba-shi, 261–8520 Japan.
a) E-mail: keiichiro.takada@sharp.co.jp
DOI: 10.1587/transinf.2023PCL0002

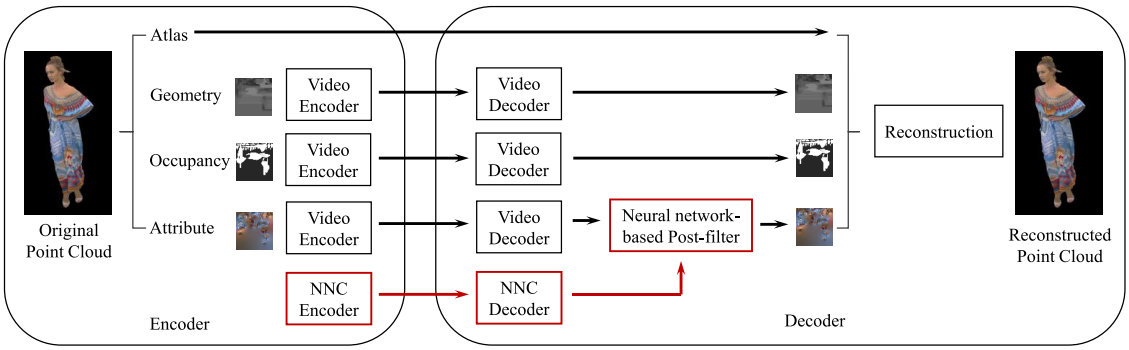

Fig. 2 Overview of proposed method: geometry, occupancy and attribute frames are encoded by video encoders and decoded by video decoders, respectively. A neural network-based post-filter is introduced to the output of the video decoder for the attribute frame.

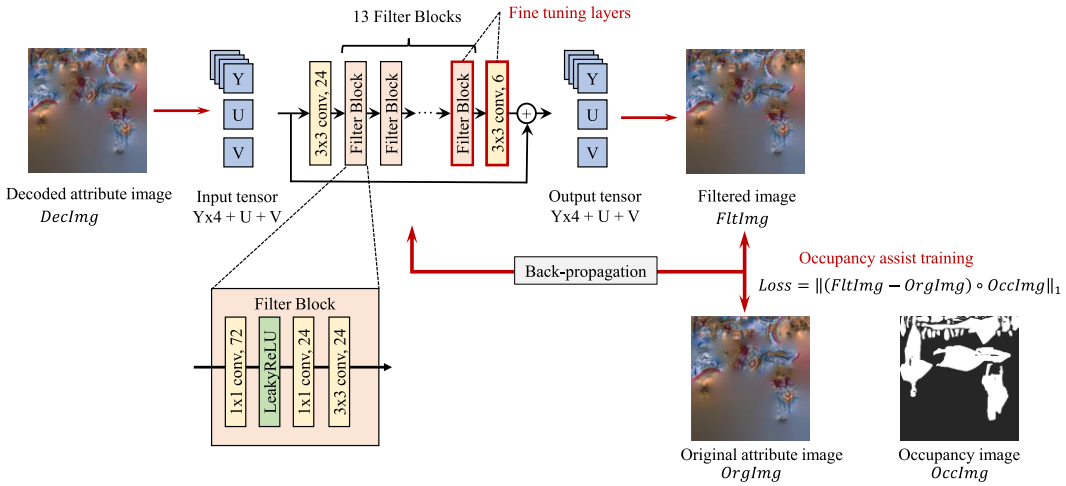

Fig. 3 Occupancy assist training: a decoded attribute image (*DecImg*) is input to the neural network model and output as a post-filtered image (*FltImg*). Then, the post-filtered image (*FltImg*) is compared to the original image (*OrgImg*), where the pixel value of the corresponding occupancy image (*OccImg*) equals one, and the derived loss value is backpropagated. The neural network model consists of an input layer, an output layer and 13 filter blocks.

(overhead) is reduced by utilizing limited layer training and incremental update functionality in NNC version 2, in which model parameters can be signalled as a difference to refer to a base model.

2.2 Occupancy Assist Training

Figure 3 shows the structure of the proposed training method, which is called Occupancy assist training. We use the occupancy image for the loss function as Eq. (1).

$$Loss = \|(FltImg - OrgImg) \circ OccImg\|_1 \quad (1)$$

where *FltImg, Orgimg*, and *OccImg* are filtered attribute, original attribute, and original occupancy images. Because the resolution of the occupancy image is lower than that of *OrgImg* and *FltImg*, and the pixel value of the image is binary, *OccImg* is converted from the image by nearest neighbour upsampling to fit the size of *FltImg*. The circle symbol in Eq. (1) denotes the element-wise product between the difference image and *OccImg* and the L1 norm is used for the

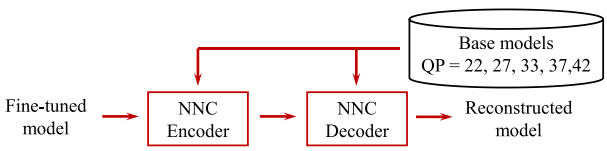

Fig. 4 Network model transmission using incremental update.

Loss. The occupancy assist training utilizes occupancy information so that the network loss is only back-propagated for regions where projected points exist.

2.3 Neural Network Model Structure

As shown Fig. 3, the input and output of the model is YUV420 whereas rearranged six channels (i.e., Y×4+U+V) format is used in input and output tensor. The model consists of a 3×3 conv input layer, 13 filter blocks and a 3×3 conv output layer. The filter block comprises 1×1 conv, Leaky ReLU, 1×1 conv and 3×3 conv. The number of

channels is from 6 to 24 at input layer, 24 to 72 at the first 1×1 conv and 72 to 24 at the second 1×1 conv in the filter blocks, and 24 to 6 at the output layer.

2.4 Model Transmission and Limited Layer Fine-Tuning

Figure 4 shows the proposed neural network model transmission using incremental update. For reference in the incremental update, base models are trained using compressed generic 2D video data. Then fine-tuned models are retrained for parameters of limited layers using locally decoded attribute frames to be transmitted. NNC encoder encodes the fine-tuned model utilizing a base model with incremental update in which only differences are to be transmitted. The decoder receives a NNC bitstream and decodes it into the reconstructed model. The base model is stored in the decoder side as out-of-band information.

Because the introduced limited layer fine-tuning ensures large part of the parameters in the fine-tuned model is the same as the base model counterpart, the transmitted bitstream size can be reduced. In this paper, the fine-tuning layers are layers in the last filter block and the 3×3 conv in the output layer.

3. Experiment

3.1 Test Condition

Experiments are performed based on V-PCC common test conditions (CTC) using V-PCC reference software, TMC2 v16 [11]. The model is compressed by a NNC reference software, NCTM (tag: INCTM-0.9) [12] with incremental update enabled. Five base models are trained using BVI-DVC [13] compressed by HM-16.20+SCM-8.8 [14] with RA configuration with five QP points (22, 27, 32, 37 and 42). Thirty-five fine-tuned models for each operation point (5 rates multiplied by 7 sequences) are trained using V-PCC CTC dataset compressed by TMC2 v16 with RA configuration with rate points (R05, R04, R03, R02 and R01) in which a base model is chosen to match the rate point (22 for R05, 27 for R04, 32 for R03, 37 for R02 and 42 for R01).

Table 1 shows the model size of the used neural network. The number of parameters is 116.5 k, and the parameter precision is float 32. The number of multiply-accumulate is 28.1 k/pixel. The size of the uncompressed base model is 446008 bytes, and the average size of fine-tuned models compressed by NNC is 6596 bytes. The neural network models are compressed into about 1.5% (one-sixty-eighth) regardless of QPs or sequences. The compressed network models were sent per 192 frames for all test sequences in this experiment. The periodic model transmission is required for scene changes and random access. (The value of 192 was chosen to balance the model's bitrate overhead and update interval. Here the overhead is around 14.0 kbps and the interval is at most every 6.4 seconds at 30 fps).

The proposed method was evaluated according to V-PCC CTC. On the CTC, the attribute data of the reconstructed point cloud are re-transformed from RGB to YUV, and PSNR values are calculated. BD-rate is derived from five points of PSNR values and total bitrates of PCC.

3.2 Result

As shown in Table 2, experimental results show BD-rate for Y, U and V compared to the anchor. The average BD-rate of the proposal, case (a) with model overhead is −3.0%, −29.3% and −22.2% for Y, U and V. The average BD-rate of comparison case (b) without NNC compressed model overhead is −3.8%, −30.3% and −23.0% for Y, U and V, showing that the model overhead incurs losses of 0.8%, 1.0% and 0.8% for Y, U and V. The relative decoding time of the proposal to the anchor is 686% (6.86 times) on CPU only.

Informal subjective evaluation was conducted comparing original, reconstructed and post-filtered images. Figure 5 (a), (b) and (c) shows rendered images of the original

Table 1 Neural network model and model size

Number of parameters	116.5k
Parameter precision	float 32
Multiply Accumulate (MAC/pixel)	28.1k
Uncompressed model size (byte)	446008
Compressed model size (byte)	6596

Table 2 BD-rate (%) for Y, U and V compared to the anchor: (a) proposed method with model overhead, (b) comparison case without model overhead, (c) comparison case without occupancy assist training and model overhead.

	(a)			(b)			(c)		
Sequence	Y	U	V	Y	U	V	Y	U	V
Loot	-1.0%	-37.3%	-32.6%	-2.6%	-38.8%	-34.4%	-1.4%	-31.0%	-27.8%
Red and black	-4.1%	-31.8%	-6.9%	-4.7%	-32.8%	-7.7%	-3.7%	-28.8%	-5.7%
Soldier	-2.4%	-31.2%	-32.7%	-3.3%	-32.5%	-33.6%	-2.3%	-23.2%	-26.3%
Queen	-1.1%	-26.1%	-17.5%	-2.3%	-27.9%	-18.7%	-1.4%	-23.5%	-15.4%
Long dress	-2.1%	-29.4%	-18.9%	-2.5%	-29.8%	-19.5%	-1.9%	-27.6%	-17.2%
Basketball player	-4.9%	-24.0%	-22.6%	-5.2%	-24.4%	-23.0%	-4.5%	-20.5%	-18.2%
Dancer	-5.5%	-25.3%	-24.0%	-5.9%	-26.1%	-24.3%	-4.9%	-19.4%	-20.4%
Average	-3.0%	-29.3%	-22.2%	-3.8%	-30.3%	-23.0%	-2.9%	-24.8%	-18.7%

Fig. 5 Rendered images of red and black, R01: (a) original image, (b) decoded image, (c) post-filtered image.

before post-filtered and after post-filtered, respectively. The rendered video was generated by mpeg-pcc-renderer [15] specified in the V-PCC test configuration. Evaluating the low-rate-point images (R01, R02 and R03), it is confirmed that the encoding artifacts, e.g. in contours and banding, are reduced without losing the features such as the hair and cloth.

3.3 Ablation Study

In addition, an ablation study for occupancy assist training was performed. Fine-tuning training without the occupancy image was tested. Otherwise, the same training procedure was used. As shown in Table 2 (c), the average BD-rate of comparison case (c) without occupancy assist training and model overhead are −2.9%, −24.8% and −18.7% for Y, U and V. Compared to the average BD-rate in Table 2 (b), occupancy image assists training shows improvements of 0.9%, 5.5% and 4.3% for Y, U and V, respectively.

4. Conclusion

This paper described a method to improve V-PCC attribute quality using a transmitted neural network model. We transmitted the neural network model employing NNC with an incremental update to achieve better coding performance. In training, occupancy image assist training and limited layer fine-tuning are applied. We confirmed the improvement of image quality is an average of 3.0%, 29.3% and 22.2% for Y, U and V, even if it included network model overhead. The subjective quality was also improved.

Acknowledgments

This work was supported by “Strategic Information and Communications R&D Promotion Programme (SCOPE)” of Ministry of Internal Affairs and Communications, Grant no. JPJ000595.

References

[1] ISO/IEC 23090-5:2021, Information technology - Coded representation of immersive media - Part 5: Visual volumetric video-based coding (V3C) and video-based point cloud compression (V-PCC), June 2021.
[2] Recommendation ITU-T H.265 | International Standard ISO/IEC 23008-2, High efficiency video coding, Aug. 2021.
[3] Recommendation ITU-T H.266 | International Standard ISO/IEC 23090-3, Versatile video coding, April 2022.
[4] T. Suzuki, T. Ikai, T. Chujoh, and N. Ito, “Latest Video Enhancement Technology for beyond H.266/VVC Video Transmission,” J. IEICE, vol.106, no.1, pp.33–38, Jan. 2023.
[5] A. Akhtar, S. Member, W. Gao, L. Li, Z. Li, W. Jia, and S. Liu, “Video-based Point Cloud Compression Artifact Removal,” IEEE Trans. Multimedia, vol.24, pp.2866–2876, June 2021. DOI: 10.1109/TMM.2021.3090148
[6] T. Chujoh, E. Sasaki, and T. Ikai, “AHG9/AHG11: Neural network based super resolution SEI,” Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-T0092, Oct. 2020.
[7] T. Suzuki, E. Sasaki, T. Chujoh, T. Ikai, and H. Watanabe, “Coded Video Transmission using Super-Resolution and MPEG-NNR Deep Learning Metadata,” Picture Coding Symposium Japan/Image Media Processing Symposium (PCSJ/IMPS 2020), P3-B-1, Nov. 2020.
[8] ISO/IEC 15938-17:2022, Information technology - Multimedia content description interface - Part 17: Compression of neural networks for multimedia content description and analysis, Aug. 2022.
[9] C. Dong, C.C. Loy, K. He, and X. Tang, “Learning a Deep Convolutional Network for Image Super-Resolution,” D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds), Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol.8692, pp.184–199, Springer, Cham, 2014. DOI: 10.1007/978-3-319-10593-2_13
[10] ISO/IEC JTC 1/SC 29/WG 7 N0038, Common Test Conditions for V3C and V-PCC, Nov. 2020.
[11] Video Point Cloud Compression - VPCC - mpeg-pcc-tmc2 test model software, https://github.com/MPEGGroup/mpeg-pcc-tmc2, 2022.
[12] Test Model of (Incremental) Compression of Neural Networks for Multimedia Content Description and Analysis, http://mpegx.int-evry.fr/software/MPEG/NNCoding/NCTM/-/tree/INCTM-0.9, 2022.
[13] D. Ma, F. Zhang, and D.R. Bull, “BVI-DVC: A Training Database for Deep Video Compression,” IEEE Trans. Multimedia, vol.24, pp.3847–3858, Sept. 2021. DOI: 10.1109/TMM.2021.3108943
[14] https://vcgit.hhi.fraunhofer.de/jvet/HM/-/tree/HM-16.20+SCM-8.8
[15] mpeg-pcc-renderer, http://mpegx.int-evry.fr/software/MPEG/PCC/mpeg-pcc-renderer, 2022.