MV-VVQA: Multi-View Learning for No-Reference Volumetric Video Quality Assessment

Yu Fan, Zicheng Zhang, Wei Sun, Xiongkuo Min, Jiaman Lin, Guangtao Zhai, and Ning Liu
Department of Electronic Engineering, *Shanghai Jiao Tong University*
fy-sky@sjtu.edu.cn

***Abstract*—Recently, volumetric video has gained growing research interest as it allows for the creation of immersive and realistic experiences by representing the full volume of 3D content. However, due to the limitation of storage space and transmission bandwidth in common applications, volumetric videos are inevitably bothered with compression and simplification distortions, which severely harms users' quality of experience (QoE). Moreover, current volumetric video quality assessment (VVQA) is mainly focused on full-reference or reduced-reference metrics, which can not be applied in the absence the reference information. Therefore, in this paper, we propose a novel deep learning based no-reference volumetric video quality assessment method based on multi-view learning. Specifically, we first project volumetric videos to 2D video sequences from various viewpoints. Then a 3D-CNN backbone is utilized to extract quality-aware features from the projected video sequences. Then a quality regression module is designed to fuse the features learned from the multiple viewpoints and jointly regress the features into quality scores. The experimental results show that our method outperforms current state-of-the-art objective volumetric video quality assessment metrics on the vsenseVVDB2 database, which validates the effectiveness of the proposed method.**

***Index Terms*—volumetric video quality assessment, no-references, multi-view, ResNet3D**

I. Introduction

Volumetric video is an emerging form of multimedia which allows viewers to perceive the video content from any viewpoint, thus providing viewers with a more immersive perceptual experience [1]. With the rapid development of computer graphics technology and depth sensors, volumetric video is easier to obtain and has been widely used in many fields, such as virtual navigation [2], immersive video conference [3], [4], sports competition [5], etc. Unlike 2D video, which is composed of 2D image frames, each frame of volumetric video consists of 3D data. The adoption of 3D point cloud as a representation of volumetric video has gained widespread acceptance due to its strong expressive ability and ease of data collection [6]. Unfortunately, the fidelity of volumetric video during transmission can be adversely affected by the limitations of network transmission and compression algorithms, leading to the degradation of the viewers' quality of experience (QoE). Therefore, there is a pressing need for an effective volumetric video quality assessment (VVQA) method to accurately evaluate the extent of such distortions.

During the last decade, many subjective VVQA studies have been carried out, during which subjects are invited to rank volumetric videos with different degrees of damage according to their personal feelings. For instance, Zerman $et\ al.$ [7], [8] collected eight volumetric video sequences and studied the subjective perception differences among different compression methods such as Darco [9], geometry-based point cloud compression (G-PCC) [10], and video-based point cloud compression (V-PCC) [10]. Cao $et\ al.$ [11] studied the effect of different bit rates and viewing distances on the subjective scoring of volumetric videos. The subjective quality evaluation method comprehensively considers the characteristics of the human visual system, and directly reflects the quality of human visual perception. However, carrying out subjective experiment is quite expensive and time-consuming, which makes it urgent to develop objective VVQA methods.

Objective quality assessment algorithms can be classified into three categories based on the involving content of reference, namely full-reference (FR), reduced-reference (RR), and no-reference (NR) methods. The MPEG Foundation introduces the p2point [12] and p2plane [12] method as an evaluation criterion for point cloud compression using the FR method. PC-MSDM [13] uses the difference in curvature between the reference point clouds and the distorted point clouds for evaluation, while PCQM [14] combines curvature with color features and establishes a linear combination parameter to obtain quality scores. GraphSIM [15] utilizes graph signal gradient to evaluate point cloud distortions. PC-SSIM [16] extracts information from geometry, normal vectors, curvature values, and colors for assessment. Viola $et\ al.$ [17] employed histogram features from geometry, luminance channel, and normal vectors to predict quality scores. 3D-NSS [18] employs color feature and geometry feature to fit parameters of Gaussian distribution to quantify distortions. Fan $et\ al.$ utilized 3D convolution networks to predict quality score [19]. These techniques constitute a range of methodologies for evaluating point cloud quality.

The aforementioned point cloud quality assessment (PCQA) methods are mainly designed for a single point cloud rather than volumetric video containing point cloud sequences. Therefore, in this paper, we propose a novel NR-VVQA method, which infers the visual quality of volumetric video from the video sequences captured from two predefined viewpoints. The viewpoints are set at the front and back side of the volumetric video's geometry center to cover sufficient quality information. Then we use the ResNet3D [20] backbone to extract features from the video sequences separately. Finally, we fuse the features from different viewpoints and adopt

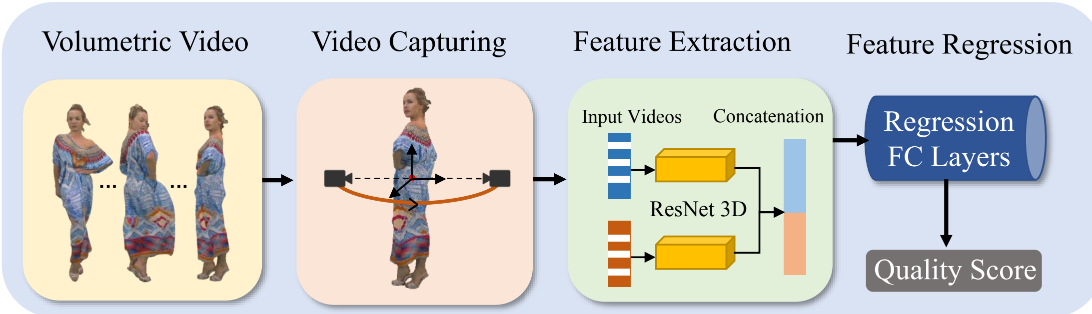

Fig. 1: The framework of the proposed method, consisting of the video capturing module, the feature extraction module and the quality score regression module.

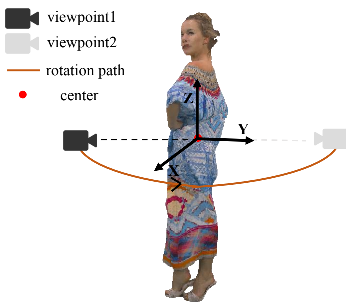

Fig. 2: Illustration of the viewpoints' positions.

fully-connected layers to predict the quality score. In the experimental section, we compare our method with current state-of-the-art FR and NR PCQA methods. To further establish the effectiveness of our methods, several video quality assessment (VQA) methods are included for comparison as well. Experimental results and statistical comparison show that our method achieves the best performance among no-reference methods on the vsenseVVDB2 database [8], which indicates the proposed method are effective for predicting the perceptual quality levels of volumetric video and can help provide useful guidelines for volumetric video compression

II. Proposed Method

The framework of our proposed method is exhibited in Fig 1, including the video capturing module, the feature extraction module and the quality score regression module.

A. Video capturing module

For the 3D volumetric video denoted as VV, we use python package Open3D [21] to generate 2D projected video sequences from two fixed viewpoints. For the i_{th} point cloud PC^i of VV, the first view point is set at the default position defined by the Open3D, which is regraded as front side of the point cloud. Then we calculate the mean center point (O_X^i, O_Y^i, O_Z^i) of the point cloud:

$$VV = \{PC^i | 1 \leq i \leq L * r\} \tag{1}$$

$$PC^i = \{pc_j^i | 1 \leq j \leq N^i\} \tag{2}$$

$$O_\alpha^i = \frac{1}{N^i} \sum_{j=1}^{N^i} pc_{j,\alpha}^i, \tag{3}$$

$$\alpha \in \{X, Y, Z\}, \tag{4}$$

where L is the length of the volumetric video VV, r is the frame rate of VV, N^i refers to the number of the points of PC^i, O_α^i refers to the X, Y, Z coordinates of the PC^i mean center, and $pc_{j,\alpha}^i$ denotes the X, Y, Z coordinates of each point in the PC^i. Then we rotate the original viewpoint 180° around the mean center (O_X^i, O_Y^i, O_Z^i) to get the second viewpoint, and Fig 2 illustrates the details of this process. The projection frame of the i_{th} point cloud are obtained by Open3D visualization function, and the captured video sequences can be derived as:

$$f_\beta^i = \mathbf{vis}(PC^i) \tag{5}$$

$$V_\beta = \{f_\beta^i | 1 \leq i \leq L * r\}, \tag{6}$$

$$\beta \in \{vp_1, vp_2\}, \tag{7}$$

where β refers to the viewpoint, **vis** is the visualization function, f_β^i refers to the projection of i_{th} point cloud corresponding to the viewpoint β, V_β consists of all projections from the certain viewpoint.

B. Feature Extraction Module

In this section we describe the process of extracting feature from the captured video sequence. The frame rate r of common volumetric video is 30, and previous research demonstrated that temporal sub-sampling can reduce computation resource consumption without sacrificing the accuracy of quality score prediction [22]. So in the training stage, we randomly select a number k between 1 and r, sample the k_{th} frame of each second, and obtain the sub-sampling video sequences $SubV_\beta$.

$$k = \mathbf{rand}(1, r) \tag{8}$$

$$SubV_\beta = \{f_\beta^{k+r*i} | 0 \leq i \leq L - 1\} \tag{9}$$

For the feature extraction of the sub-sampling video sequence, we utilize the ResNet3D network as the backbone. As 3D convolution do temporal convolution and spatial convolution simultaneously, ResNet3D network can extract feature involving both temporal and spatial information. For each viewpoint, we employ an independent ResNet3D network for its feature extraction, and the process can be concluded as:

$$F_\beta = R3D_\beta(SubV_\beta) \tag{10}$$
$$\beta \in \{vp_1, vp_2\}, \tag{11}$$

where $R3D$ is the ResNet3D network, F_β indicates the extracted features of ResNet3D network from different viewpoints.

C. Feature Regression Module

The feature regression module takes the extracted ResNet3D feature as input, and outputs the overall quality score. To fuse the feature from different viewpoints, the feature vector from different viewpoints are concatenated together and two fully-connected layers with 1024 neurons and 256 neurons are utilized. The final score Q_p are calculated as:

$$F_{in} = F_{vp_1} \oplus F_{vp_2} \tag{12}$$
$$Q_p = FC(F_{in}) \tag{13}$$

where $\oplus$ means the concatenation operation, F_{in} are the input feature of regression module and FC are the fully-connected layers.

The loss function for the network is mean squared error (MSE) loss:

$$Loss = ||\boldsymbol{Q}_p - \boldsymbol{Q}_{gt}||_2^2 \tag{14}$$

where Q_{gt} is the ground truth mean opinion score (MOS).

III. Experiment

A. Database

We conduct experiments on the vsenseVVDB2 database [8] with the volumetric video consisting of point clouds and the volumetric videos consisting of mesh are excluded. The database contains 8 reference volumetric video sequences and each volumetric video lasts for ten seconds with frames rate 30, which indicates that 300 = 30×10 point clouds are included. Three MPEG standard compression algorithms,including G-PCC with region-adaptive hierarchical transform (RAHT), V-PCC with all-intra (AI) mode, V-PCC with random-acess (RA) mode, are utilized to compress these volumetric video at different bit-rate, which generates 128=8×16 compressed point cloud volumetric video in total.

B. Experiment Setup

In this section, we explain the details of our experiment. Due to the scale of current volumetric video database, we do a 8-fold cross validation to maximize the utilization of available data. Each time we select seven reference volumetric videos, employ their distorted videos for network training, and leave the remained volumetric video's distorted versions as the test set. After 8 rounds, all groups of volumetric videos are tested, and we record the average performance as the results. We use the Adam optimizer [23] with initiate learning rate 1e-5, the batch size is set to 4 and the number of epoch is 30. The input video frames are resized to 480×480, and randomly cropped into 448× 448 patches.

To evaluate the correlation between predicted quality score and MOS, we employ four widely-used correlation evaluation metrics. Root mean square error (RMSE) denotes the error gap between predicted score and MOS. Spearman Rank Correlation Coefficient (SRCC) and Kendall's Rank Correlation Coefficient (KRCC) evaluate the degree of monotonicity. Pearson Linear Correlation Coefficient (PLCC) measures the linear correlation. The value range for SRCC, KRCC, PLCC is [-1,1], and a higher value means better performance.

TABLE I: Performance results on the vsenseVVDB2 databases.

Index	Type	Methods	SRCC	PLCC	KRCC	RMSE
A	FR	p2point(RMS)	0.6726	0.7908	0.4950	10.4964
B		p2point(Haus)	0.6055	0.6748	0.4623	12.6551
C		p2plane(RMS)	0.5434	0.5685	0.3849	17.1490
D		p2plane(Haus)	0.5356	0.6454	0.3982	13.0985
E		psnr-Y	0.6229	0.7389	0.4752	11.5536
F		PC-SSIM	0.6853	0.8224	0.5476	9.7556
G		PCQM	0.7540	0.8767	0.5694	8.2486
H		GraphSIM	0.7730	0.8854	**0.6111**	7.9682
I	NR	3D-NSS	**0.7793**	**0.8972**	0.6061	**7.5702**
J		BRISQUE	0.3126	0.3567	0.1845	20.8431
K		VSFA	0.5919	0.7861	0.4583	10.2463
L		StairVQA	0.7414	0.7721	0.6367	11.5326
M		Proposed	**0.8648**	**0.9007**	**0.7199**	**7.9271**

C. Compared Methods

As there is no specific designed VVQA methods, we compare our proposed method with several PCQA methods. For these PCQA methods, we record the quality score of each single point cloud frame, and apply the average pooling to obtain the quality score of the volumetric video. To extend the range of comparison, we also take some famous NR VQA methods for comparison. All compared methods are as follows:

- FR methods: FR metrics include p2point [12], p2plane [24], psnr-Y [25], PCQM [14], GraphSIM [15], and PC-SSIM [16]. Note p2point and p2plane is evaluated with different distance criteria: root mean squared (RMS) distance and Hausdorff (Haus) distance .
- NR methods: NR metrics consist of 3D-NSS [18], BRISQUE [26], VSFA [27], and StairVQA [28]. Note that Brisque, VSFA and StairVQA are NR-VQA methods and utilize the same setup as our proposed method.

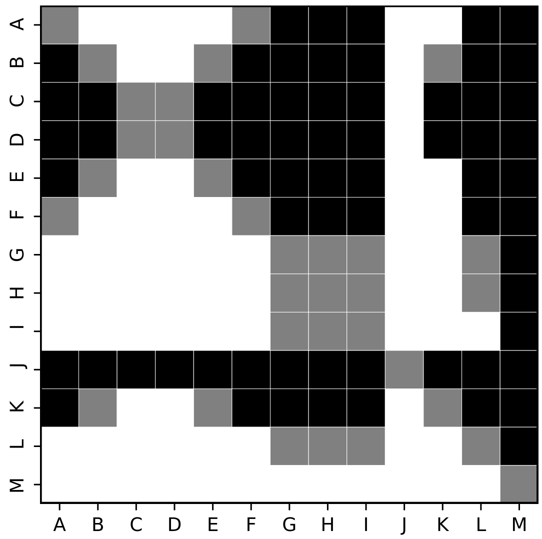

Fig. 3: Statistical significance test results on vsenseVVDB2 database. A white/black block indicates that the row model is statistically better/worse than the column model. A gray block indicates that the row and column models are statistically indistinguishable. A-N are model indices given in Table I.

TABLE II: Ablation study of the viewpoints.

viewpoint	SRCC	PLCC	KRCC	RMSE
vp1	0.8268	0.8974	0.7000	8.1598
vp2	0.8489	0.8928	0.7166	8.0353
vp1+vp2	0.8648	0.9007	0.7199	7.9271

D. Results

The results of each method are reported in Table I, and the best performance is marked in red, and sub-optimal result is marked in blue, where we can find our proposed method achieves the best performance in SRCC, PLCC, KRCC and sub-optimal performance in RMSE. The underlying factors contributing to the performance are explained as follows. Compared with the PCQA-based methods, we leveraged the ResNet3D network to extract spatial-aware and temporal-aware features. Spatial features are useful for the evaluation of blocking artifact caused by compression, and temporal feature are employed to aggregate features from different frames to the overall quality score, rather than average pooling. As for the VQA-based methods, we adopt multi-viewpoints and make full use of videos captured from different angles.

To further validate whether the results from different methods are statistically better or worse, a t-test on the SRCC values is adopted as recommended by [29], and the results are demonstrated in Fig 3, where a white/black block denotes the row model is statistically better/worse than the column model, and a gray block denotes the row model has similar performance with the column model. It's clear that our method is statistically better than current PCQA methods or VQA methods, which demonstrates the effectiveness of our proposed method.

An ablation study is also carried out to ascertain the impact of using videos from different viewpoints. The results are listed in Table II, where $vp1$ means only using the captured video from viewpoint1, so as $vp2$. It is clear that the combination of different viewpoints improves the performance of quality score prediction.

IV. Conclusion

In this paper we propose a novel framework to deal with the VVQA task. To better utilize the free viewpoint characteristics of volumetric video, we employ two different viewpoints to capture videos from source volumetric videos, and the ResNet3D backbone are applied to extract both temporal and spatial aware features with 3D convolution. The extracted feature from different viewpoints are fused and utilized to predict the final quality score. The experimental results and statistical T-test demonstrated that our proposed method outperforms current state-of-the-art full-reference and no-reference PCQA metrics on the vsenseVVDB2 databases, which reflects the effectiveness of the proposed method.

References

[1] Aljoscha Smolic, "3d video and free viewpoint video—from capture to display," *Pattern recognition*, vol. 44, no. 9, pp. 1958–1968, 2011.

[2] Olgierd Stankiewicz, Marek Domański, Adrian Dziembowski, Adam Grzelka, Dawid Mieloch, and Jarosław Samelak, "A free-viewpoint television system for horizontal virtual navigation," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2182–2195, 2018.

[3] Simon NB Gunkel, Hans M Stokking, Rick Hindriks, and Tom de Koninck, "Vr conferencing: communicating and collaborating in photo-realistic social immersive environments.," *International Journal of Virtual Reality*, 2019.

[4] Yizhong Zhang, Jiaolong Yang, Zhen Liu, Ruicheng Wang, Guojun Chen, Xin Tong, and Baining Guo, "Virtualcube: An immersive 3d video communication system," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 5, pp. 2146–2156, 2022.

[5] Jun Chen, Ryosuke Watanabe, Keisuke Nonaka, Tomoaki Konno, Hiroshi Sankoh, and Sei Naito, "Fast free-viewpoint video synthesis algorithm for sports scenes," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3209–3215.

[6] Shishir Subramanyam, Jie Li, Irene Viola, and Pablo Cesar, "Comparing the quality of highly realistic digital humans in 3dof and 6dof: A volumetric video case study," in *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2020, pp. 127–136.

[7] Emin Zerman, Pan Gao, Cagri Ozcinar, and Aljosa Smolic, "Subjective and objective quality assessment for volumetric video compression," in *IS&T Electronic Imaging, Image Quality and System Performance XVI*, 2019.

[8] Emin Zerman, Cagri Ozcinar, Pan Gao, and Aljosa Smolic, "Textured mesh vs coloured point cloud: A subjective study for volumetric video compression," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[9] Google, "Draco: 3d data compression," https://github.com/google/draco/, 2017.

[10] Sebastian Schwarz, Marius Preda, Vittorio Baroncini, Madhukar Budagavi, Pablo Cesar, Philip A Chou, Robert A Cohen, Maja Krivokuća, Sébastien Lasserre, Zhu Li, et al., "Emerging mpeg standards for point cloud compression," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, pp. 133–148, 2018.

[11] Keming Cao, Yi Xu, and Pamela Cosman, "Visual quality of compressed mesh and point cloud sequences," *IEEE Access*, vol. 8, pp. 171203–171217, 2020.

[12] Rufael Mekuria, Zhu Li, Christian Tulvan, and Phil Chou, "Evaluation criteria for point cloud compression," *ISO/IEC MPEG*, , no. 16332, 2016.

[13] Gabriel Meynet, Julie Digne, and Guillaume Lavoué, "Pc-msdm: A quality metric for 3d point clouds," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–3.

[14] Gabriel Meynet, Yana Nehmé, Julie Digne, and Guillaume Lavoué, "Pcqm: A full-reference quality metric for colored 3d point clouds," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[15] Qi Yang, Zhan Ma, Yiling Xu, Zhu Li, and Jun Sun, "Inferring point cloud quality via graph similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3015–3029, 2020.

[16] Evangelos Alexiou and Touradj Ebrahimi, "Towards a point cloud structural similarity metric," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[17] Irene Viola and Pablo Cesar, "A reduced reference metric for visual quality evaluation of point cloud contents," *IEEE Signal Processing Letters*, vol. 27, pp. 1660–1664, 2020.

[18] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, and Guangtao Zhai, "No-reference quality assessment for 3d colored point cloud and mesh models," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7618–7631, 2022.

[19] Yu Fan, Zicheng Zhang, Wei Sun, Xiongkuo Min, Ning Liu, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai, "A no-reference quality assessment metric for point cloud based on captured video sequences," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2022, pp. 1–5.

[20] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[21] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.

[22] Ali Ak, Emin Zerman, Suiyi Ling, Patrick Le Callet, and Aljosa Smolic, "The effect of temporal sub-sampling on the accuracy of volumetric video quality assessment," in *2021 Picture Coding Symposium (PCS)*. IEEE, 2021, pp. 1–5.

[23] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[24] Dong Tian, Hideaki Ochimizu, Chen Feng, Robert Cohen, and Anthony Vetro, "Geometric distortion metrics for point cloud compression," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 3460–3464.

[25] Alain Horé and Djemel Ziou, "Image quality metrics: Psnr vs. ssim," *2010 20th International Conference on Pattern Recognition*, pp. 2366–2369, 2010.

[26] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[27] Dingquan Li, Tingting Jiang, and Ming Jiang, "Quality assessment of in-the-wild videos," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2351–2359.

[28] Wei Sun, Tao Wang, Xiongkuo Min, Fuwang Yi, and Guangtao Zhai, "Deep learning based full-reference and no-reference quality assessment models for compressed ugc videos," in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2021, pp. 1–6.

[29] David J Sheskin, *Handbook of parametric and nonparametric statistical procedures*, Chapman and Hall/CRC, 2003.