

GEOMETRY-AWARE VIDEO QUALITY ASSESSMENT FOR DYNAMIC DIGITAL HUMAN

Zicheng Zhang^{1,2}, Yingjie Zhou^{1,2}, Wei Sun^{1,2}, Xionghuo Min^{1,2}, and Guangtao Zhai^{1,2,3}

¹Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China

²Peng Cheng Laboratory, China

³MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China

zzc1998@sjtu.edu.cn

ABSTRACT

Dynamic Digital Humans (DDHs) are 3D digital models that are animated using predefined motions and are inevitably bothered by noise/shift during the generation process and compression distortion during the transmission process, which needs to be perceptually evaluated. Usually, DDHs are displayed as 2D rendered animation videos and it is natural to adapt video quality assessment (VQA) methods to DDH quality assessment (DDH-QA) tasks. However, the VQA methods are highly dependent on viewpoints and less sensitive to geometry-based distortions. Therefore, in this paper, we propose a novel no-reference (NR) geometry-aware video quality assessment method for DDH-QA challenge. Geometry characteristics are described by the statistical parameters estimated from the DDHs' geometry attribute distributions. Spatial and temporal features are acquired from the rendered videos. Finally, all kinds of features are integrated and regressed into quality values. Experimental results show that the proposed method achieves state-of-the-art performance on the DDH-QA database.

Index Terms— Dynamic digital human, video quality assessment, no-reference, geometry-aware

1. INTRODUCTION

With the increasing popularity of digital humans in various applications, such as virtual reality, gaming, and telecommunication, the quality of dynamic digital humans (DDHs) has become a crucial factor in providing a realistic and engaging experience. To provide useful guidelines for compression algorithms and improve the Quality of Experience (QoE) of viewers, it is necessary to carry out objective quality assessment methods to predict the quality values for DDHs. Considering that the DDHs are usually rendered into 2D animation videos for exhibition [1], it is reasonable to trans-

fer video quality assessment (VQA) methods to DDH quality assessment (DDH-QA) tasks. During the last decade, large amounts of effort have been dedicated to pushing forward the development of VQA. Early full-reference (FR) VQA methods typically use IQA methods, such as PSNR and SSIM [2], to compute the quality difference between reference and distorted frames. Similar to FR-VQA methods, some no-reference (NR) VQA methods compute each frame's quality level using NR image quality assessment (IQA) methods, such as BRISQUE [3] and NIQE [4]. To incorporate spatial and temporal features, handcrafted-based methods have been proposed, such as VIIDEO [5], V-BLIINDS [6], TLVQM [7], and VIDEVAL [8]. Deep neural networks (DNNs) have also been employed, such as VSFA [9], RAPIQUE [10], SimpVQA [11], and FAST-VQA [12, 13].

However, the rendered videos are variant to the viewpoints and the 2D media are not sensitive to the 3D model distortions [14, 15], which indicates simply employing VQA methods is far from enough for DDH-QA task. Therefore, in this paper, we propose a novel NR geometry-aware VQA method to deal with DDH-QA issues. Specifically, geometry attributes including dihedral angle and curvature are computed for the 3D geometry mesh of the digital humans. Then statistical parameters are estimated from the geometry attribute distributions to quantify the geometry distortions such as geometry noise and compression. Afterward, the rendered 2D videos are split into clips for spatial and temporal feature extraction. The first frame of each clip is used for spatial feature extraction with a 2D-CNN backbone while each whole clip is utilized for temporal feature extraction with a fixed pretrained 3D-CNN backbone. The spatial features can help identify the texture distortions like blur and color noise while the temporal features can assist detect motion distortions including motion blur and motion unnaturalness. Later, the geometry, spatial, and temporal features are fused with concatenation and regressed into quality scores with fully-connected (FC) layers. The experimental results show that the proposed method outperforms all the comparing methods on the DDH-QA [16] database.

This work was supported in part by NSFC (No.62225112, No.61831015), the Fundamental Research Funds for the Central Universities, National Key R&D Program of China 2021YFE0206700, Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), and STCSM 22DZ2229005.

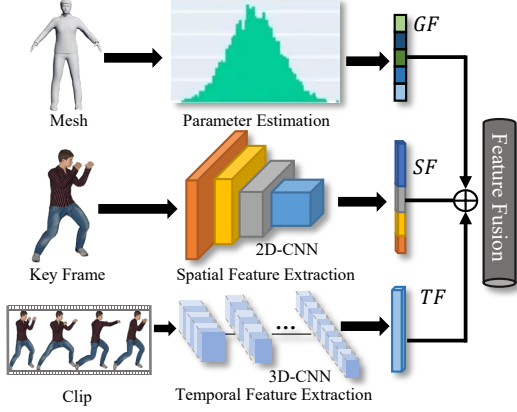


Fig. 1. The framework of the proposed method, where the geometry, spatial, and temporal features are extracted from the geometry mesh, key frame, and clip respectively. Then the features are fused with concatenation.

2. PROPOSED METHOD

Given a DDH model, we first derive the static digital human geometric mesh \mathcal{M} . Afterward, the DDH model is rendered into an animation video sequence \mathcal{V} from a perceptually selected viewpoint, which can cover the major quality information of the DDH model. Both \mathcal{M} and \mathcal{V} are directly provided in the DDH-QA [16] database.

2.1. Geometry Feature Extraction

It has been shown in previous works [17] that the geometry characteristics are effective for describing the quality-aware local patterns of the 3D models. A mesh is typically defined as a collection of vertices, edges, and faces, then we define the geometry mesh for the DDH as:

$$\mathcal{M} = (\mathbf{Vt}, \mathbf{Eg}, \mathbf{Fc}), \quad (1)$$

where \mathbf{Vt} , \mathbf{Eg} , and \mathbf{Fc} represent the sets of vertices, edges, and faces respectively.

2.1.1. Dihedral Angle

The dihedral angle is the angle between the normals of two adjacent faces, which has been regarded as an effective indicator for quantifying the quality of mesh simplification and compression algorithms [18]. We can calculate the dihedral angle between two adjacent faces in a mesh by the dot product of corresponding normal vectors:

$$\cos \theta_i = \frac{\mathbf{n}_{i1} \cdot \mathbf{n}_{i2}}{\|\mathbf{n}_{i1}\| \|\mathbf{n}_{i2}\|}, \quad (2)$$

where θ_i denotes the dihedral angle of i -th edge Eg_i , \mathbf{n}_{i1} and \mathbf{n}_{i2} represent the normal vectors of the two adjacent faces whose coedge is Eg_i . For each edge in the set \mathbf{Eg} , its corresponding dihedral angle is computed, which finally generates an array of dihedral angles θ .

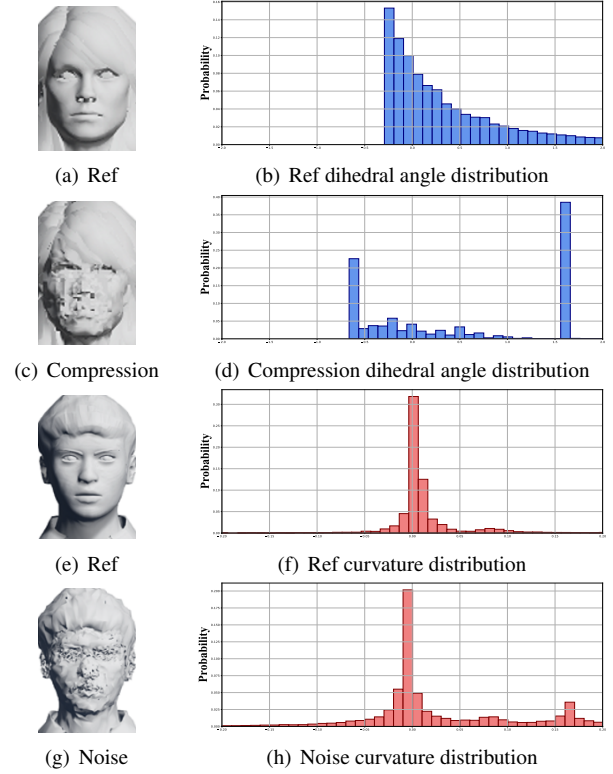


Fig. 2. Reference and distortion examples for the static mesh model along with the corresponding normalized probability distributions for dihedral angle and curvature respectively. The reference distributions can be greatly altered by the presence of distortions. And the estimated statistical parameters are capable of reflecting the perceptual loss from the distribution appearance, as proven in [17].

2.1.2. Gaussian Curvature

Curvature is commonly used for characterizing the features of a surface, such as describing smoothness or roughness, which makes it quite sensitive to structural distortions and thus enables it to assist in describing the visual quality of 3D models [17, 19]. To quantify structural damage for the mesh, we choose Gaussian curvature as the corresponding feature operator:

$$G_m = \frac{(2\pi - \sum_n \theta_{mn})}{A_m}, \quad (3)$$

where G_m is the Gaussian curvature for the m -th vertex Vt_m , θ_{mn} is the n -th angle between the two adjacent edges at vertex Vt_m , and A_i indicates the area of the Voronoi cell of vertex Vt_m . For each vertex in the set \mathbf{Vt} , its corresponding Gaussian curvature is computed, which finally generates an array of Gaussian curvature \mathbf{G} .

2.1.3. Stastical Parameters Estimation

The mean, variance, and entropy are employed as the basic statistical parameters. We further choose the generalized

Gaussian distribution (GGD) [3], the general asymmetric generalized Gaussian distribution (AGGD) [3], and the Gamma distribution to estimate quality parameters from the normalized dihedral angle array $\hat{\theta}$ and the normalized Gaussian curvature array \hat{G} . It has been proven that the appearance of such feature distributions can be altered by various types of distortions, which can be reflected by the estimated statistical parameters [17, 19]. Therefore, these parameters are effective for measuring the visual fidelity of 3D models in the presence of distortion, which can be calculated as:

$$\begin{aligned}\mathcal{X} &\sim \text{Basic}(\mu, \sigma^2, E), \\ \hat{\mathcal{X}} &\sim \text{GGD}(\alpha_1, \beta_1), \\ \hat{\mathcal{X}} &\sim \text{AGGD}(\eta, v, \sigma_l^2, \sigma_r^2), \\ \hat{\mathcal{X}} &\sim \text{Gamma}(\alpha_2, \beta_2),\end{aligned}\quad (4)$$

where \mathcal{X} and $\hat{\mathcal{X}}$ represent the distributions and normalized distributions for dihedral angle and curvature arrays, (μ, σ^2, E) parameters in the $\text{Basic}(\cdot)$ function stand for (mean, variance, entropy) respectively. More specifically, the GGD parameters estimation can be obtained as:

$$\text{GGD}(x; \alpha_1, \beta_1^2) = \frac{\alpha_1}{2\beta_1\Gamma(1/\alpha_1)} \exp\left(-\left(\frac{|x|}{\beta_1}\right)^{\alpha_1}\right), \quad (5)$$

where $\beta_1 = \sigma\sqrt{\frac{\Gamma(1/\alpha_1)}{\Gamma(3/\alpha_1)}}$, $\Gamma(\alpha_1) = \int_0^\infty t^{\alpha_1-1}e^{-t}dt$, $\alpha_1 > 0$ is the gamma function, and the two estimated parameters (α_1, β_1^2) indicate the shape and variance of the distribution. The AGGD parameters estimation can be derived as:

$$\begin{aligned}\text{AGGD}(x; \eta, v, \sigma_l^2, \sigma_r^2) &= \\ &\begin{cases} \frac{v}{(\beta_l + \beta_r)\Gamma(\frac{1}{v})} \exp\left(-\left(\frac{-x}{\beta_l}\right)^v\right), & x < 0, \\ \frac{v}{(\beta_l + \beta_r)\Gamma(\frac{1}{v})} \exp\left(-\left(\frac{x}{\beta_r}\right)^v\right), & x \geq 0, \end{cases}\end{aligned}\quad (6)$$

where η represents the β_r and β_l difference while $\beta_l = \sigma_l\sqrt{\Gamma(\frac{1}{v})/\Gamma(\frac{3}{v})}$ and $\beta_r = \sigma_r\sqrt{\Gamma(\frac{1}{v})/\Gamma(\frac{3}{v})}$, σ_l^2 and σ_r^2 characterize the spread extent of the distribution on the left and right sides, v determines the shape of the distribution. The shape-rate Gamma distribution is formulated as:

$$\text{Gamma}(x; \alpha_2, \beta_2) = \frac{\beta_2^{\alpha_2} x^{\alpha_2-1} e^{-\beta_2 x}}{\Gamma(\alpha_2)} x > 0, \quad (7)$$

where α_2 and β_2 stands for the shape and rate parameters and $\alpha_2, \beta_2 > 0$. In all, a total of $2 \times (3+2+4+2) = 22$ statistical parameters are obtained for describing the geometry perceptual quality for a single DDH and we refer to these features as GF .

2.2. Video Feature Extraction

Given a rendered animation video whose number of frames and frame rate is n_f and r_f , we split the video into $\frac{n_f}{r_f}$ clips for feature extraction and each clip lasts for 1s.

2.2.1. Spatial Feature Extraction

The spatial features can directly assist the model to identify the existence and extent of common distortions such as blur and noise. Additionally, considering the hierarchical visual perception process, we employ the multi-scale features extracted from a 2D-CNN backbone to incorporate the quality-aware information from low-level to high-level. For the i -th clip C_i , the first frame is selected as the key frame for spatial feature extraction:

$$\begin{aligned}SF_i &= \alpha_1 \oplus \alpha_2 \oplus \alpha_3 \oplus \dots \oplus \alpha_{N_L}, \\ \alpha_j &= \text{GAP}(L_j(F_i)), j \in \{1, 2, 3, \dots, N_L\},\end{aligned}\quad (8)$$

where SF_i denotes the extracted spatial features from the key frame of the i -th clip, \oplus represents the concatenation operation, $\text{GAP}(\cdot)$ stands for the global average pooling operation, $L_j(F_i)$ indicates the feature maps obtained from j -th layer of the 2D-CNN backbone, α_j denotes the corresponding average pooled features, and N_L is the number of the layers for the 2D-CNN.

2.2.2. Temporal Feature Extraction

DDHs can be bothered by motion-based distortions such as motion unnaturalness and model clipping. Therefore, to capture the motion-based quality-aware features, we utilize a pre-trained 3D-CNN backbone for temporal feature extraction:

$$TF_i = \mathcal{T}(C_i), \quad (9)$$

where TF_i represents the extracted temporal features from the i -th clip C_i and \mathcal{T} indicates the feature extraction operation of the pretrained 3D-CNN backbone.

2.3. Feature Fusion & Quality Regression

With the geometry features and video features extracted above, we conduct the clip-level feature fusion by concatenation:

$$F_i = GF \oplus SF_i \oplus TF_i, \quad (10)$$

where GF represents the geometry features for the DDH, SF_i and TF_i indicate the spatial and temporal features extracted from C_i , and F_i is the final fused features for C_i . Then two-stage fully-connected layers are employed to regress the clip-level features into quality values:

$$Q_i = FC(F_i), \quad (11)$$

where Q_i stands for the predicted quality score for clip C_i and the final quality can be computed via average pooling:

$$\mathcal{Q} = \frac{1}{N_C} \sum_{i=1}^{N_C} Q_i, \quad (12)$$

where \mathcal{Q} is the final quality score for the DDH and N_C indicates the number of used clips.

Table 1. Performance results on the DDH-QA database. Best in **RED** and second in **BLUE**.

Ref.	Model	SRCC \uparrow	PLCC \uparrow	KRCC \uparrow	RMSE \downarrow
FR	PSNR	0.4308	0.5458	0.3114	0.9013
	SSIM	0.5408	0.6057	0.3920	0.8559
NR	BRISQUE	0.3664	0.4011	0.2568	1.0067
	NIQE	0.0923	0.2489	0.0748	1.0418
	VIIDEO	0.1219	0.1829	0.0732	1.0740
	V-BLIINDS	0.4807	0.4936	0.3424	0.9564
	TLVQM	0.2515	0.2824	0.1729	1.0480
	VIDEVAL	0.5218	0.3470	0.1622	1.0246
	VSFA	0.5406	0.5708	0.3858	0.9657
	RAPIQUE	0.1815	0.2368	0.1246	1.0614
	SimpVQA	0.7444	0.7498	0.5452	0.7228
	FAST-VQA	0.5262	0.5382	0.3657	1.0499
Proposed	0.8004	0.7956	0.6028	0.6343	

3. EXPERIMENT

3.1. Experimental Setup & Implementation Details

The DDH-QA [16] database is employed for validation, which provides 800 degraded DDHs with both model-based and motion-based distortions. The DDH videos are with varying durations (2s~8s) and are divided into 1s clips. We uniformly employ 6 clips from each video with cyclic sampling. Specifically, if a video has less than 6 clips, the existing clips are expanded with cyclic sampling until 6 clips are selected. For videos lasting for more than 6 seconds, the first 6 clips are used. The ResNet50 [20] is employed as the spatial feature extractor and patches with a resolution of $448 \times 448 \times 3$ are cropped as input. The SlowFast R50 [21] is utilized as the temporal feature extractor and the clips are resized to 224×224 for both training and testing. The ResNet50 is initialized with a pre-trained model on the ImageNet database [22] and fine-tuned during the training phase. While the SlowFast R50 is frozen, with pre-trained model weights on the Kinetics 400 database [23]. The Adam optimizer [24] is utilized, with an initial learning rate of $4e-6$. The default number of epochs and batch size are set as 30 and 4. The mean squared error (MSE) is used as the loss function.

The 5-fold cross-validation strategy is employed. In this strategy, the 10 motion groups are split into 5 folds, with each fold containing 2 groups of motion. Four folds are utilized as training sets, while the remaining fold is used as the testing set. This process is repeated 5 times, ensuring that every fold is used as the testing set. The final experimental results are obtained by recording the average performance. Furthermore, for methods that do not require training, we apply them to the same testing sets and report their average performance.

3.2. Benchmark Competitors & Criteria

To validate the animated videos in the DDH-QA database, several video quality assessment (VQA) methods are utilized. The FR methods, such as PSNR and SSIM [2], operate on the frame level of the DDH videos. The NR methods include handcrafted-based methods, such as BRISQUE [3], NIQE

Table 2. Experimental performance of the ablation study, where GF, SF, and TF indicate the geometry features, spatial features, and temporal features respectively.

Feature	SRCC \uparrow	PLCC \uparrow	KRCC \uparrow	RMSE \downarrow
GF+SF	0.7771	0.7762	0.5896	0.6888
GF+TF	0.4312	0.4903	0.2961	0.9393
SF+TF	0.7786	0.7731	0.5860	0.6702
All	0.8004	0.7956	0.6028	0.6343

Table 3. SRCC Experimental performance corresponding to the used number of clips. Since the videos last for 2s~8s, we test the proposed method with numbers of clips from 2~8.

Num	2	3	4	5	6	7	8
SRCC	0.6914	0.7150	0.7501	0.7711	0.8004	0.7850	0.7745

[4], VIIDEO [5], V-BLIINDS [6], TLVQM [7], and VIDEVAL [8], as well as DNN-based methods, such as VSFA [9], RAPIQUE [10], SimpleVQA [11], and FAST-VQA [12].

Four mainstream consistency evaluation criteria are utilized to compare the correlation between the predicted scores and MOSs, which include Spearman Rank Correlation Coefficient (SRCC), Kendall’s Rank Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Squared Error (RMSE).

3.3. Performance Discussion

The experimental performance is exhibited in Table 1, from which we can see that the proposed method outperforms all the compared methods and surpass the second-place method SimpVQA by about 7.5% in terms of SRCC, which indicates the effectiveness of the proposed method for evaluating the perceptual quality of DDHs. To further investigate the contributions of different types of features, we conduct the ablation study and the results are shown in Table 2. With closer inspection, we can find that using all three types of features achieves the best performance, which reveals that GF, SF, and TF make contributions to the final results. Moreover, we test the influence of utilizing varying numbers of clips and the results are illustrated in Table 3. It can be seen that when using smaller than 6 clips, the performance can be improved by the increasing number of clips. However, using 7 or 8 clips can cause performance drops due to redundancy and over-fitting.

4. CONCLUSION

In conclusion, this paper proposes a novel no-reference geometry-aware video quality assessment method for Dynamic Digital Humans. By leveraging statistical parameters estimated from DDHs’ geometry attribute distributions and spatio-temporal features acquired from rendered videos, the proposed method achieves state-of-the-art performance on the DDH-QA database. The approach offers an effective solution to the DDH quality assessment (DDH-QA) tasks, which can be useful for improving the visual quality of DDHs and enhancing the user experience in various applications.

5. REFERENCES

- [1] Zicheng Zhang, Wei Sun, Yucheng Zhu, Xiongkuo Min, Wei Wu, Ying Chen, and Guangtao Zhai, “Treating point cloud as moving camera videos: A no-reference quality assessment metric,” *arXiv preprint arXiv:2208.14085*, 2022.
- [2] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.
- [3] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE TIP*, vol. 21, no. 12, pp. 4695–4708, 2012.
- [4] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.
- [5] Anish Mittal, Michele A Saad, and Alan C Bovik, “A completely blind video integrity oracle,” *IEEE TIP*, vol. 25, no. 1, pp. 289–300, 2015.
- [6] Michele A Saad, Alan C Bovik, and Christophe Charrier, “Blind prediction of natural video quality,” *IEEE TIP*, vol. 23, no. 3, pp. 1352–1365, 2014.
- [7] Jari Korhonen, “Two-level approach for no-reference consumer video quality assessment,” *IEEE TIP*, vol. 28, no. 12, pp. 5923–5938, 2019.
- [8] Zhengzhong Tu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, “Ugc-vqa: Benchmarking blind video quality assessment for user generated content,” *IEEE TIP*, vol. 30, pp. 4449–4464, 2021.
- [9] Dingquan Li, Tingting Jiang, and Ming Jiang, “Quality assessment of in-the-wild videos,” in *ACM MM*, 2019, pp. 2351–2359.
- [10] Zhengzhong Tu, Xiangxu Yu, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik, “Rapique: Rapid and accurate video quality prediction of user generated content,” *IEEE OJSP*, vol. 2, pp. 425–440, 2021.
- [11] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai, “A deep learning based no-reference quality assessment model for ugc videos,” in *ACM MM*, 2022, pp. 856–865.
- [12] Haoning Wu, Chaofeng Chen, Jingwen Hou, Liang Liao, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin, “Fast-vqa: Efficient end-to-end video quality assessment with fragment sampling,” in *ECCV*. 2022, pp. 538–554, Springer.
- [13] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin, “Neighbourhood representative sampling for efficient end-to-end video quality assessment,” *arXiv preprint arXiv:2210.05357*, 2022.
- [14] Zicheng Zhang, Wei Sun, Xiongkuo Min, Quan Zhou, Jun He, Qiyuan Wang, and Guangtao Zhai, “Mm-pcqa: Multi-modal learning for no-reference point cloud quality assessment,” *IJCAI*, 2023.
- [15] Zicheng Zhang, Wei Sun, Houning Wu, Yingjie Zhou, Chunyi Li, Xiongkuo Min, Guangtao Zhai, and Weisi Lin, “Gms-3dqa: Projection-based grid mini-patch sampling for 3d model quality assessment,” *arXiv preprint arXiv:2306.05658*, 2023.
- [16] Zicheng Zhang, Yingjie Zhou, Wei Sun, Wei Lu, Xiongkuo Min, Yu Wang, and Guangtao Zhai, “Ddh-qa: A dynamic digital humans quality assessment database,” *IEEE ICME*, 2023.
- [17] 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 TCSVT*, 2022.
- [18] N. Mukherjee, “A hybrid, variational 3d smoother for orphaned shell meshes,” in *IMR*, 2002.
- [19] Zicheng Zhang, Wei Sun, Xiongkuo Min, Tao Wang, Wei Lu, Wenhan Zhu, and Guangtao Zhai, “A no-reference visual quality metric for 3d color meshes,” in *ICMEW*. IEEE, 2021, pp. 1–6.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *IEEE/CVF CVPR*, 2016, pp. 770–778.
- [21] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He, “Slowfast networks for video recognition,” in *IEEE/CVF CVPR*, 2019, pp. 6202–6211.
- [22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *IEEE/CVF CVPR*. Ieee, 2009, pp. 248–255.
- [23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.
- [24] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *ICLR*, 2014.