CAP: An Advanced No-Reference Quality Assessment Method for AI-Generated 3D Meshes

Yingjie Zhou, Farong Wen, Zicheng Zhang, Yanwei Jiang, Jun Jia, Xiaohong Liu, Xiongkuo Min, Guangtao Zhai
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, China
zyj2000@sjtu.edu.cn

***Abstract*—The advent of generative AI has revolutionized 3D content design, significantly enhancing modelers' efficiency. However, the quality of generated 3D content, particularly Generated Meshes (GMs), remains a critical concern. GMs pose unique challenges for quality assessment due to their complex geometry, detailed texture mapping, and distortions that differ from traditional meshes. Existing methods fail to address these GM-specific issues. To tackle this gap, we introduce a novel no-reference quality assessment method, CAP, which integrates CT-Slice, prompt Alignment, and Projections. CAP employs a six-face projection to capture external features and a CT-like slicing approach to extract internal quality features. Additionally, it leverages Contrastive Language-Image Pre-Training (CLIP) to measure the alignment between projection embeddings and prompts as a key quality indicator. Experimental results demonstrate that CAP effectively evaluates GM quality by combining internal, external, and alignment features. The code for this work has been open-sourced in https://github.com/zyj-2000/CAP.**

***Index Terms*—AIGC, projection-based, 3D quality assessment, no-reference, CLIP**

I. Introduction

3D content has garnered significant attention due to its immersive depth and three-dimensional (3D) visual effects, making it a cornerstone of virtual reality (VR) applications [1]–[4]. However, traditional 3D modeling processes are time-intensive and laborious for designers. Even with advanced sensing technologies capable of capturing geometric and color information from real-world objects, substantial manual refinement is often required, and the equipment itself is costly. These challenges have hindered the widespread adoption of 3D content and slowed the advancement of related technologies. The advent of generative AI has transformed 3D content creation, introducing methods tailored to Colored Point Clouds, Textured Meshes, Neural Radiation Fields (NeRF) [5], and 3D Gaussian Splatting (3DGS) [6], thereby expanding designers' creative options. Despite these innovations, quality assurance of 3D AI-generated content (3DGC) remains a pressing concern due to limited human oversight and insufficient feedback mechanisms. Among various 3DGC types, Generated Meshes (GMs) stand out for their accurate representation of object shapes and surface details through intricate data structures, making quality issues particularly pronounced. Thus, developing a robust and reliable quality assessment framework for GMs is crucial to advancing 3D Artificial Intelligence Generated Content (AIGC) and improving the end-user visual experience.

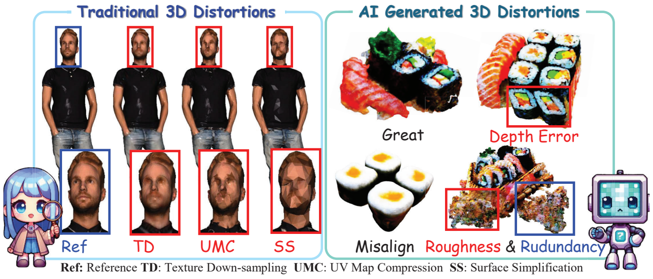
Traditional 3D Distortions
AI Generated 3D Distortions
Ref
TD
UMC
SS
Great
Depth Error
Misalign
Roughness & Rudundancy
Ref: Reference TD: Texture Down-sampling UMC: UV Map Compression SS: Surface Simplification

Fig. 1. Visualization of different types of distortion suffered by the conventional meshes and the generated meshes.

Numerous representative 3D Quality Assessment (3DQA) methods [7]–[13] have been developed, offering valuable insights for evaluating GMs. However, most existing 3DQA methods are tailored for captured 3D content or computer-simulated distortions, overlooking the unique distortion types specific to GMs. As illustrated in Fig. 1, traditional meshes typically experience distortions due to pre-processing during transmission, as well as simplifications and compression for storage optimization. These distortions are visually intuitive and easily identifiable. In contrast, GMs primarily suffer distortions arising from the 3D generation algorithms and the prompts used, which are often more complex and less apparent. These distortions necessitate consideration of alignment with the prompts, making traditional mesh quality assessment (MQA) methods unsuitable for direct application to GMs. To address this gap, we propose a novel quality assessment method for GMs, CAP. Unlike existing projection-based 3DQA methods [10], [11], [11], [14], [15], which focus solely on external mesh features, CAP integrates internal structural analysis and prompt alignment as critical quality indicators. Specifically, CAP employs six-face projections to comprehensively capture the external appearance of GMs. It further examines internal structural features through CT-like slicing along multiple X-Y planes. Additionally, it utilizes Contrastive Language-Image Pre-Training (CLIP) [16] to measure the alignment between six projection embeddings and prompt embeddings. Experimental results on the 3DGCQA dataset [17] demonstrate that CAP significantly outperforms

This work was supported in part by National Natural Science Foundation of China (623B2073, 62101326, 62225112).

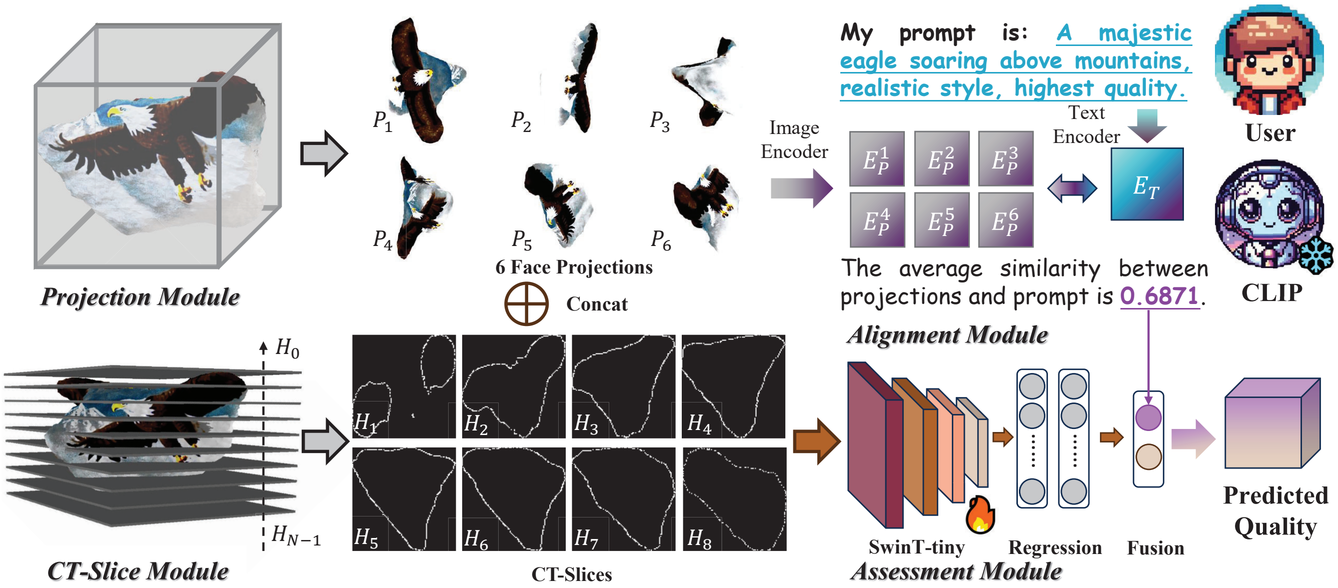

Fig. 2. Framework of proposed CAP method.

existing objective quality assessment methods, offering an effective solution for evaluating GMs' quality.

II. Related Work

A. AI Generated Content Quality Assessment

The advent of generative AI has profoundly influenced various aspects of production and daily life, particularly within the art and design industries [18]–[23]. While generative models have significantly enhanced practitioners' efficiency, the quality of AIGC has become a critical area of focus. In this regard, Li *et al.* pioneered the assessment of AI-generated images by introducing the AGIQA dataset [24], [25], which was later extended by studies addressing AI-generated video quality. Despite advancements in quality assessment for AI-generated images (AGI) and videos (AGV) [26], the quality evaluation of generated 3D content has received comparatively less attention due to the inherent complexity of 3D generation. Recognizing this gap, Zhou *et al.* [17] created the first dedicated dataset for 3D content quality assessment, 3DGCQA, which includes 313 Generated Meshes (GMs). Although the dataset provides valuable resources for advancing the field, Zhou *et al.* did not propose an effective quality assessment method. To address this limitation, this paper introduces a novel and reliable quality assessment method specifically designed for GMs, building upon the foundation laid by existing datasets and research.

B. 3D Quality Assessment

Advancements in media technology have brought increasing attention to 3D content, which offers depth of information and immersive, interactive experiences surpassing those of 2D media such as images and videos. However, the complex data structure of 3D content introduces a broader range of quality issues. To address these challenges, several foundational 3D quality assessment (3DQA) datasets [27]–[31] have been developed, providing essential data for evaluating 3D content quality. Classical 3DQA methods are typically categorized based on the type of 3D data: point cloud quality assessment (PCQA) [10], [11], [15] and mesh quality assessment (MQA) [12], [14], [27]. In point clouds, common distortions include video-based point cloud compression (VPCC) and geometry-based point cloud compression (GPCC), geometric and color noise, and downsampling. In meshes, distortions generally involve geometric and texture compression, quantization, and noise. To address these issues, various 3DQA methods have been proposed, broadly classified into model-based [7]–[9] and projection-based [10], [11], [15] approaches. Among these, projection-based methods have gained prominence due to their simplicity, computational efficiency, and strong performance. However, the emergence of generative 3D models has introduced new challenges for 3DQA. Traditional distortions are no longer dominant in GMs, which instead exhibit unique distortions. Zhou *et al.* [17] have identified several such distortions, including multi-faceted duplicates, depth errors, surface roughness, geometric redundancy, and geometric deletions, which are not adequately addressed by existing 3DQA methods. These new challenges necessitate novel approaches to effectively evaluate the quality of GMs.

III. Proposed Method

Given the substantial differences in quality assessment requirements between Generated Meshes (GMs) and traditional meshes, we propose a novel quality assessment method, CAP, as illustrated in Fig. 2. The CAP framework comprises four distinct modules. The Projection Module, CT-Slice Module, and Alignment Module are designed to assess distinct aspects

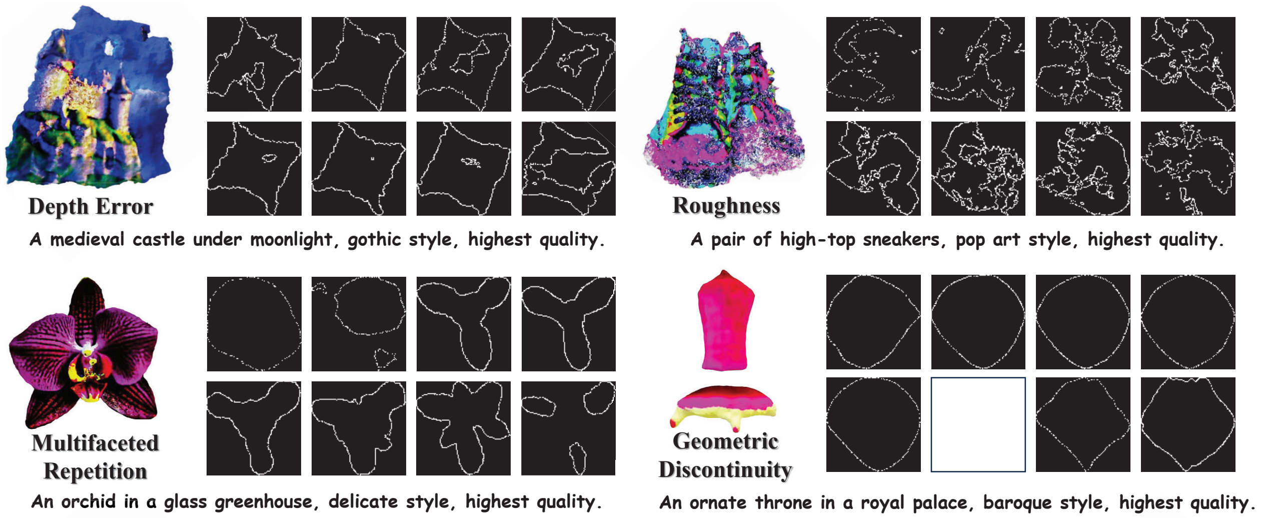
Depth Error
A medieval castle under moonlight, gothic style, highest quality.
Roughness
A pair of high-top sneakers, pop art style, highest quality.
Multifaceted Repetition
An orchid in a glass greenhouse, delicate style, highest quality.
Geometric Discontinuity
An ornate throne in a royal palace, baroque style, highest quality.

Fig. 3. Slices of Generated Mesh with different distortion types.

of GM quality, focusing on external quality information, internal structural features, and prompt alignment, respectively. The Assessment Module integrates these insights by performing feature extraction, regression, and fusion to produce the final predicted quality score of the GM.

A. Projection Module

Projection is a common preprocessing technique in 3DQA, enabling the transformation of complex 3D models into 2D images using virtual cameras [32]. This approach effectively reduces dimensionality and computational overhead. Among projection schemes, frontal projection and six-sided projection are the two prevailing methods. Given the visually complex distortions often present in GMs, the six-face projection employed in CAP provides a more comprehensive visualization of a GM's external features. To formalize the process, a GM is defined as:

$$GM = (V, N, E), \tag{1}$$

where $V = \{v_1, v_2, \ldots, v_i\}$, $N = \{n_1, n_2, \ldots, n_j\}$, $E = \{e_1, e_2, \ldots, e_k\}$ are the set of i vertices, j normal vectors and k edges in GM respectively. The geometric center of the GM is calculated as:

$$V_c = \frac{1}{|V|} \sum_{v \in V} v, \tag{2}$$

where $|V|$ denotes the total number of vertices and V_c denotes the center coordinates of GM. Since there is insufficient prior information to determine the frontal surface of each GM, we adopt the initial viewpoint of the virtual camera as the frontal surface and a Cartesian coordinate system is established with V_c as the origin. To ensure comprehensive external visualization, the virtual camera is positioned along the positive and negative directions of the X, Y and Z axes at a fixed distance r. This configuration enables the generation of six distinct projections $P_i (i = 1, 2, \ldots, 6)$, which effectively capture the GM's external visual information.

B. CT-Slice Module

Although the six-face projections provide sufficient information to describe the external appearance of GMs, they inevitably lose depth and structural details. Many GM distortions exploit this limitation, creating high-quality renderings from specific viewpoints while masking overall quality deficiencies. To address this potential for deceptive visualizations, it is crucial to perform CT slicing on the GM to capture its internal structural information. In the established Cartesian coordinate system, we first identify the highest and lowest vertices, denoted as $\hat{v}_h$ and $\hat{v}_l$, respectively. From these, we compute the height H and the slicing interval I of the GM:

$$I = \frac{H}{N-1} = \frac{\hat{v}_h - \hat{v}_l}{N-1}, \tag{3}$$

where N is the number of slices. The slices are then defined along the Z-axis from high to low, labeled from H_0 to H_{N-1}. Notably, the first and last CT slices pass through only a few vertices at the highest and lowest points (including vertices of equal height), which do not provide sufficient structural information. Besides, they are already represented by the six-face projections. Therefore, the number of valid slices is $N-2$. To illustrate the relationship between CT slices and GM quality, we perform slicing on typical GMs with various distortion types, as shown in Fig. 3. The results reveal that CT slices effectively capture the GM's geometric structure, with distinct distortions manifesting differently in the slices. Specifically, GMs with depth errors exhibit concave surfaces in the slices, while rough geometric structures create closed regions with smaller, irregular areas. Furthermore, multifaceted repetitive distortions are fully captured in the CT slices, while geometric discontinuities may result in blank slices.

C. Alignment Module

A key distinction between traditional MQA and the evaluation of GMs is that the latter requires an objective quality assessment that not only considers the external visual and internal structural quality but also evaluates the consistency of the generated content with the provided prompts. To address this, we propose the Alignment Module, which computes the similarity between the six-face projection embedding and the prompt embedding using CLIP [16]. Specifically, the process begins by encoding the projection embedding E_P^i for each projection $P_i (i = 1, 2, \dots, 6)$ using a pre-trained Vision Transformer [33]. Simultaneously, the prompt is encoded using Transformer [34] to obtain its corresponding embedding E_T. The similarity between the six projection embeddings and the prompt embedding is then calculated via a dot-product operation. The average of these similarity values is taken as the measure of consistency between the GM and prompt, serving as a key feature for quality assessment.

D. Assessment Module

Using the Projection Module and the CT-Slice Module, we obtain six projections that capture the external visual information of the GM and $N - 2$ slices that represent the internal geometric structure. Given the superior performance of the Swin Transformer (Swin-T) [35] in various computer vision tasks, we employ Swin-T tiny for further feature extraction from both the six projections and the $N - 2$ slices. Subsequently, two fully connected (FC) layers are used as quality regressors to predict the quality of the GM. To compute the final quality score, we adaptively weigh the average similarity from the Alignment Module and the predicted quality from the FC layers using an additional FC layer with two neurons. For model training and parameter updates, we utilize the Mean Squared Error (MSE) as the loss function.

IV. Experiments

A. Experiment Setups

To comprehensively evaluate and analyze the performance of the proposed CAP method, we conduct both performance tests and ablation experiments. For the dataset, we utilize the 3DGCQA dataset, the only dataset specifically designed for quality assessment of 3D-generated content. To demonstrate the effectiveness of CAP, we select 11 representative objective quality assessment methods for comparison under consistent experimental conditions. Notably, four methods including DBCNN [36], StairIQA [37], ViT-MQA [14] and Dual-PCQA [15] require additional training, while Q-Align [38], based on a multimodal large language model (MLLM), is tested only in a zero-shot setting. The remaining methods rely on manually extracted features. All methods employ the same six-face projection strategy, and the average quality score is used as the predicted GM quality. For CAP training, the number of slices is set to $N = 10$. We use the Adam optimizer [39] with default settings of 30 epochs and a learning rate of 5e-5. The two FC layers of the regressor have 768 and 64 neurons, respectively. We follow the experimental setup of [17], employing five-fold cross-validation to evaluate CAP's performance on the 3DGCQA dataset. Care is taken to ensure no content overlap between folds, and the average performance across all five folds is recorded for each method.

TABLE I
Performance comparison of different methods on 3DGCQA database. Best in **RED**, second in **BLUE**.

Type	Method	SRCC↑	PLCC↑	KRCC↑	RMSE↓
Hand-crafted Based	BRISQUE [40]	0.2091	0.3347	0.1444	0.7414
	CPBD [41]	0.2099	0.4797	0.1335	0.7217
	IL-NIQE [42]	0.1481	0.1573	0.1131	0.6600
	NFERM [43]	0.2797	0.4062	0.1999	0.7222
	NFSDM [44]	0.3189	0.4468	0.2235	0.6935
	NIQE [45]	0.2079	0.2594	0.1413	0.8050
Deep-learning Based	DBCNN [36]	0.5381	0.5147	0.3946	0.4700
	StairIQA [37]	0.3813	0.4566	0.2653	0.5802
	ViT-MQA [14]	0.3517	0.3724	0.2609	0.8780
	Dual-PCQA [15]	**0.6583**	**0.6578**	**0.4718**	**0.4549**
	Q-Align [38]	0.0746	0.0764	0.0498	0.8311
	CAP (Ours)	**0.7098**	**0.7566**	**0.5386**	**0.4245**

B. Experiment Criteria

To quantify the performance of each method on 3DGCQA dataset, we utilize four widely adopted metrics in quality assessment: Spearman Rank Order Correlation Coefficient (SRCC), Pearson Linear Correlation Coefficient (PLCC), Kendall Rank Order Correlation Coefficient (KRCC), and Root Mean Square Error (RMSE). The first three metrics range from 0 to 1, with values closer to 1 indicating better performance of the assessment method. Conversely, RMSE evaluates the accuracy of the predicted quality, with lower values closer to 0 reflecting higher predictive accuracy.

C. Performance Analysis

The performance of various objective quality assessment methods on the 3DGCQA dataset is summarized in Table I, from which several conclusions can be drawn: 1) The proposed CAP method achieves state-of-the-art (SOTA) performance, outperforming all other methods with a substantial margin (+5% SRCC). This demonstrates the effectiveness of the proposed approach; 2) Methods relying on manually extracted features generally perform poorly, indicating that quality assessment algorithms designed for natural scene images (NSIs) are unsuitable for generative content. This disparity arises from the significant differences in prior distributions between generative content and NSIs. Additionally, traditional quality assessment algorithms, optimized for classical distortions such as blur and noise, fail to address the more complex and nuanced distortions characteristic of generative content; 3) While existing deep learning based methods exhibit better performance, their primary limitation is the exclusive focus on external visual information of GMs. These methods neglect both the distortion characteristics unique to GMs and the alignment of generated content with the provided prompts, which are critical factors in the quality assessment of generative meshes.

TABLE II
ABLATION STUDY RESULTS ON 3DGCQA DATABASE, WHERE C, A, P DENOTE CT-SLICE, ALIGNMENT AND PROJECTION MODULES, RESPECTIVELY. BEST IN **RED**, SECOND IN **BLUE**.

Model	SRCC↑	PLCC↑	KRCC↑	RMSE↓
C	0.4917	0.4923	0.3615	0.5972
A	0.3371	0.4093	0.2402	0.5781
P	0.5048	0.5689	0.3590	0.5580
$A + P$	0.5558	0.5896	0.3939	0.5241
$C + P$	**0.6695**	**0.6611**	**0.4907**	**0.4470**
$C + A$	0.5477	0.5683	0.4017	0.5399
$C + A + P$	**0.7098**	**0.7566**	**0.5386**	**0.4245**

D. Ablation Experiments

To evaluate the effectiveness of each module within the proposed CAP method, ablation experiments are conducted, and the results are presented in Table II. Analysis of these results yields several key insights: 1) Each module positively contributes to the overall performance of CAP, affirming the importance and utility of these components; 2) The CT-Slice and Projection modules demonstrate comparable performance in the ablation experiments, highlighting that the internal structural information of GMs is as critical as their external visual features for objective quality assessment; 3) When combined with the data from Table I, the alignment between GM and prompt surpasses traditional manual feature extraction methods. This finding underscores the significant influence of prompt alignment on the overall quality assessment of GMs.

V. CONCLUSION

While 3D Generated Content (3DGC) offers significant benefits by simplifying 3D modeling and enhancing design efficiency, its quality remains a critical concern due to limited human supervision and a lack of effective control mechanisms. Among various types of 3DGC, Generative Meshes (GMs) are particularly vulnerable to distortions due to their complex and large-scale data structures. These distortions severely impact user experience and visual quality. Unlike traditional 3D distortions, which primarily result from trade-offs in communication transmission and storage constraints, distortions in GMs are more visually deceptive and challenging to detect. Consequently, many existing objective 3D quality assessment methods fail to effectively evaluate GM quality. To address these challenges, this paper introduces a novel objective quality assessment method for GMs, termed CAP. CAP integrates three key modules: CT-Slice, Alignment, and Projection Modules. This approach captures both the internal structural features and external visual characteristics of GMs while incorporating the degree of alignment between the GM and its associated prompts as a critical quality metric. Experimental results demonstrate that CAP significantly outperforms existing methods in assessing GM quality, providing a robust and reliable quality indicator.

REFERENCES

[1] Yingjie Zhou, Yaodong Chen, Kaiyue Bi, Lian Xiong, and Hui Liu, "An implementation of multimodal fusion system for intelligent digital human generation," *arXiv preprint arXiv:2310.20251*, 2023.

[2] Yu Zhou, Yanjing Sun, Leida Li, et al., "Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network," *IEEE TCSVT*, vol. 32, no. 4, pp. 1767–1777, 2021.

[3] Yu Zhou, Weikang Gong, Yanjing Sun, et al., "Pyramid feature aggregation for hierarchical quality prediction of stitched panoramic images," *IEEE TMM*, vol. 25, pp. 4177–4186, 2022.

[4] Yu Zhou, Weikang Gong, Yanjing Sun, et al., "Quality assessment for stitched panoramic images via patch registration and bidimensional feature aggregation," *IEEE TMM*, 2023.

[5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, et al., "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, et al., "3d gaussian splatting for real-time radiance field rendering.," *ACM TOG*, vol. 42, no. 4, pp. 139–1, 2023.

[7] Eric M Torlig, Evangelos Alexiou, Tiago A Fonseca, et al., "A novel methodology for quality assessment of voxelized point clouds," in *Applications of Digital Image Processing XLI*, 2018, vol. 10752, pp. 174–190.

[8] Paolo Cignoni, Claudio Rocchini, and Roberto Scopigno, "Metro: measuring error on simplified surfaces," in *Computer graphics forum*. Wiley Online Library, 1998, vol. 17, pp. 167–174.

[9] Rufael Mekuria, Zhu Li, Christian Tulvan, et al., "Evaluation criteria for point cloud compression," 2016.

[10] Zicheng Zhang, Wei Sun, Yingjie Zhou, et al., "Eep-3dqa: Efficient and effective projection-based 3d model quality assessment," in *ICME*, 2023, pp. 2483–2488.

[11] Zicheng Zhang, Wei Sun, Haoning Wu, et al., "Gms-3dqa: Projection-based grid mini-patch sampling for 3d model quality assessment," *ACM TOMM*, vol. 20, no. 6, pp. 1–19, 2024.

[12] Zicheng Zhang, Yingjie Zhou, Chunyi Li, et al., "A reduced-reference quality assessment metric for textured mesh digital humans," in *ICASSP*. IEEE, 2024, pp. 2965–2969.

[13] Zicheng Zhang, Haoning Wu, Yingjie Zhou, et al., "Lmm-pcqa: Assisting point cloud quality assessment with lmm," in *ACM MM*, 2024, pp. 7783–7792.

[14] Yingjie Zhou, Zicheng Zhang, Wei Sun, et al., "A no-reference quality assessment method for digital human head," in *ICIP*. IEEE, 2023, pp. 36–40.

[15] Zicheng Zhang, Yingjie Zhou, Wei Sun, et al., "Simple baselines for projection-based full-reference and no-reference point cloud quality assessment," *arXiv preprint arXiv:2310.17147*, 2023.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, et al., "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[17] Yingjie Zhou, Zicheng Zhang, Farong Wen, et al., "3dgcqa: A quality assessment database for 3d ai-generated contents," in *ICASSP*. IEEE, 2025, pp. 1–5.

[18] Zicheng Zhang, Wei Sun, Yingjie Zhou, et al., "Subjective and objective quality assessment for in-the-wild computer graphics images," *ACM MM*, vol. 20, no. 4, pp. 1–22, 2023.

[19] Yingjie Zhou, Zicheng Zhang, Wei Sun, et al., "Thqa: A perceptual quality assessment database for talking heads," in *ICIP*, 2024, pp. 15–21.

[20] Zicheng Zhang, Haoning Wu, Chunyi Li, et al., "A-bench: Are lmms masters at evaluating ai-generated images?," *arXiv preprint arXiv:2406.03070*, 2024.

[21] Zicheng Zhang, Yingjie Zhou, Chunyi Li, et al., "Quality assessment in the era of large models: A survey," *ACM TOMM*, 2024.

[22] Xiongkuo Min, Huiyu Duan, Wei Sun, et al., "Perceptual video quality assessment: A survey," *Science China Information Sciences*, vol. 67, no. 11, pp. 211301, 2024.

[23] Yingjie Zhou, Zicheng Zhang, Jiezhang Cao, et al., "Memo-bench: A multiple benchmark for text-to-image and multimodal large language models on human emotion analysis," *arXiv preprint arXiv:2411.11235*, 2024.

[24] Chunyi Li, Zicheng Zhang, Haoning Wu, et al., "Agiqa-3k: An open database for ai-generated image quality assessment," *TCSVT*, 2023.

[25] Chunyi Li, Tengchuan Kou, Yixuan Gao, et al., “Aigiqa-20k: A large database for ai-generated image quality assessment,” *arXiv preprint arXiv:2404.03407*, 2024.

[26] Zhichao Zhang, Wei Sun, Xinyue Li, et al., “Human-activity agv quality assessment: A benchmark dataset and an objective evaluation metric,” *arXiv preprint arXiv:2411.16619*, 2024.

[27] Zicheng Zhang, Yingjie Zhou, Wei Sun, et al., “Perceptual quality assessment for digital human heads,” in *ICASSP*. IEEE, 2023, pp. 1–5.

[28] Zicheng Zhang, Yingjie Zhou, Wei Sun, et al., “Ddh-qa: A dynamic digital humans quality assessment database,” in *ICME*. IEEE, 2023, pp. 2519–2524.

[29] Zicheng Zhang, Wei Sun, Yingjie Zhou, et al., “Advancing zero-shot digital human quality assessment through text-prompted evaluation,” *arXiv preprint arXiv:2307.02808*, 2023.

[30] Zicheng Zhang, Yingjie Zhou, Long Teng, et al., “Quality-of-experience evaluation for digital twins in 6g network environments,” *IEEE TBC*, 2024.

[31] Yingjie Zhou, Zicheng Zhang, Wei Sun, et al., “Subjective and objective quality-of-experience assessment for 3d talking heads,” in *ACM MM*, 2024, pp. 6033–6042.

[32] Yingjie Zhou, Zicheng Zhang, Wei Sun, et al., “Perceptual quality assessment for point clouds: A survey,” *ZTE COMMUNICATIONS*, vol. 21, no. 4, 2023.

[33] Alexey Dosovitskiy, “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[34] A Vaswani, “Attention is all you need,” *NIPS*, 2017.

[35] Ze Liu, Yutong Lin, Yue Cao, et al., “Swin transformer: Hierarchical vision transformer using shifted windows,” in *ICCV*, 2021, pp. 10012–10022.

[36] Weixia Zhang, Kede Ma, Jia Yan, et al., “Blind image quality assessment using a deep bilinear convolutional neural network,” *IEEE TCSVT*, vol. 30, no. 1, pp. 36–47, 2020.

[37] Wei Sun, Xiongkuo Min, Guangtao Zhai, et al., “Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training,” *arXiv preprint arXiv:2105.14550*, 2021.

[38] Haoning Wu, Zicheng Zhang, Weixia Zhang, et al., “Q-align: Teaching lmms for visual scoring via discrete text-defined levels,” in *ICML*.

[39] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.

[40] 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.

[41] Niranjan D Narvekar and Lina J Karam, “A no-reference image blur metric based on the cumulative probability of blur detection (cpbd),” *IEEE TIP*, vol. 20, no. 9, pp. 2678–2683, 2011.

[42] L. Zhang, L. Zhang, and A. C. Bovik, “A feature-enriched completely blind image quality evaluator,” *IEEE TIP*, vol. 24, no. 8, pp. 2579–2591, 2015.

[43] Ke Gu, Guangtao Zhai, Xiaokang Yang, et al., “Using free energy principle for blind image quality assessment,” *IEEE TMM*, vol. 17, no. 1, pp. 50–63, 2014.

[44] Ke Gu, Guangtao Zhai, Xiaokang Yang, et al., “No-reference image quality assessment metric by combining free energy theory and structural degradation model,” in *IEEE ICME*, 2013, pp. 1–6.

[45] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE SPL*, vol. 20, no. 3, pp. 209–212, 2013.