SI23DCQA: Perceptual Quality Assessment of Single Image-to-3D Content

Kang Fu*, Huiyu Duan*, Zicheng Zhang, Xiaohong Liu, Xiongkuo Min†, Jia Wang, and Guangtao Zhai†
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

***Abstract*—In recent years, significant efforts have been dedicated to advancing 3D content generation. However, existing quality assessment research predominantly focuses on evaluating Text-to-3D Content (T23DC) while ignoring Single Image-to-3D Content (SI23DC). In this paper, we establish the first Single Image-to-3D Content Quality Assessment (SI23DCQA) database to comprehensively study the perceptual quality of SI23DCs. The database contains 1500 SI23DCs, which are generated by 5 common SI23DC algorithms from 300 images including realistic images, AI generated images, and model rendered images. Afterward, we carry out a well-designed subjective experiment to collect subjective quality ratings for SI23DCs from three perspectives including overall, color, and shape. Additionally, a benchmark experiment is conducted with the state-of-the-art no reference image quality assessment (NR-IQA), no reference video quality assessment (NR-VQA), and no reference 3D quality assessment (NR-3DQA) and the experimental results show that current quality assessment methods are limited in evaluating the perceptual loss of SI23DCs. The database is released on https://github.com/ZedFu/SI23DCQA.**

***Index Terms*—artificial intelligence generated content, quality assessment database, 3D quality assessment**

I. Introduction

With the advancement of generation model, artificial intelligence generated content (AIGC) algorithms provide a new way to generate contents including text, image, video, 3D model, *etc.*, and have been used to many areas such as search sites, games, films, and entertainment, *etc.* At the same time, many studies have focused on artificial intelligence generated content quality assessment (AIGCQA) [1]–[5] and have established many corresponding subjective quality assessment database. For the AI generated 3D content, it can be divided into two categories, including T23DC and SI23DC. However, current quality assessment studies only focus on evaluating T23DC [5] while ignoring SI23DC. Compared with the T23DC, SI23DC generally generates higher quality and more detailed 3D contents, since the input image can provide more information than prompt text. Meanwhile, the subjective quality assessment manner of SI23DC is quite different from T23DC. The former needs to pay attention to the correspondence of color and shape between the generated 3D contents and the input images, while the latter needs to focus on the alignment between the generated 3D contents and the prompt texts. SI23DCQA is both significant and promising, yet there is currently no existing work dedicated to addressing it.

This work was supported in part by the National Natural Science Foundation of China under Grant (62401365, 62225112, 62271312, 62132006, 623B2073). * Equal contribution. † Corresponding authors (Email: {minxiongkuo,zhaiguangtao}@sjtu.edu.cn)

In order to address this problem, we establish the first SI23DC database to tackle the mentioned challenge. First, we collect 300 images of three types: realistic images, AI generated images and model rendered images. Then we utilize 5 common SI23DC algorithms to generate 1500 3D contents. Afterward, we carry out a well-designed subjective experiment to collect the perceptual quality scores from three perspectives including color, shape and overall. of the generated 3D contents. Finally, we conduct a benchmark experiment on the proposed database with state-of-the-art NR-IQA, NR-VQA and NR-3DQA methods, which shows that current quality assessment methods are not effective for predicting the visual quality levels of SI23DCs. Our contributions can be summarized as followed:

- **To the best of our knowledge, we establish the first Single Image-to-3D content quality assessment database**, which contains 1,500 3D contents generated by 5 SI23DC algorithms from 300 images of three types: realistic images, AI generated images and model rendered images.
- We carry out a subjective study to collect three types perceptual quality scores (overall, color, and shape) of the generated 3D contents. A total of $94,500 = 21 \times 3 \times 1,500$ quality ratings are gathered.
- We also conduct a benchmark experiment to exhibit the performance of the existing state-of-the-art NR-IQA, NR-VQA and NR-3DQA methods.

II. Related Work

A. Single Image-to-3D Content Generation

Single Image-to-3D Content, the task of generating 3D content from a single 2D image, is a challenging and meaningful problem in the computer version community and is essential for numerous applications, such as 3D content creation, AR/VR as well as robotic manipulation and navigation [11], [12]. There have been extensive efforts to address this problem. Early approaches learn 3D priors from 3D synthetic data or real scans. Since 3D data can be represented in various formats, some approaches focus on generating point clouds, meshes or voxels, while others concentrate on generating implicit representations, such as SDFs and NeRF. However these methods are category-specific generation and produce barely satisfactory results. In recent years, the vision-language models and 2D generation models have trained on Internet-scale image datasets and learned human-like visual concepts,

TABLE I
SUMMARY OF THE EXISTING AIGCQA DATABASES AND PROPOSED SI23DCQA DATABASE. THE NUMBERS IN PARENTHESES OF SCORE TYPE REPRESENT THE DIMENSIONS OF THE SUBJECTIVE EXPERIMENTAL ANNOTATIONS.

Type	Dataset	Contents	Prompts	Models	Annotators	Ratings	Score type
Text-To-Image	AGIQA-1K [1]	1,080	540	2	22	23,760	MOS
	AGIQA-3K [2]	2,982	497	6	21	125,244	MOS(2)
	AGIQA-20K [6]	20,000	20,000	15	21	420,000	MOS
	AIGCIQA2023 [3]	2,400	100	6	28	201,600	MOS(3)
	AIGCOIQA2024 [7]	300	25	5	20	18,000	MOS(3)
Text-To-Video	Chivileva's [8]	1,005	201	5	24	48,240	MOS(2)
	EvalCrafter [9]	3,500	700	7	3	73,500	MOS(5)
	FETV [10]	2,476	619	3	3	11,142	MOS(2)
	T2VQA-DB [4]	1,000	1,000	9	27	27,000	MOS
Text-To-3D	3DGCQA [5]	313	50	7	40	25,040	MOS(2)
Single Image-To-3D	Ours	1500	300	5	21	94,500	MOS(3)

which can possess powerful priors about our 3D world. Therefore, some recent methods [11], [13] first utilize view-conditioned 2D generation model which can generate multi-view images though inputted single image, and then feed these multi-view images to 3D reconstruction model to generate 3D contents. In the same time, with the release of large-scale 3D datasets [14], some approaches [12], [15], [16] adopt the encoder-decoder framework, which first use pretrained vision model to encoder the input image as a middle tensor and convert it to 3D triplane representation by a large transformer decoder.

B. AI Generated Content Quality Assessment

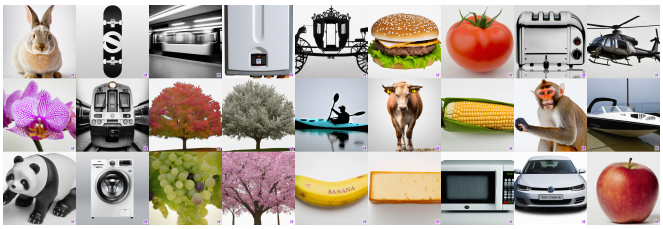
In recent years, diffusion model has achieved great success in generation tasks and can generate very realistic contents, which leads to an exponential increase of AIGC in daily lives. In the same time, many researchers have established lots of databases and proposed many methods to evaluate the quality of generated content. Some work [1]–[3], [6], [7] focus on AI generated image quality assessment and other work [4], [8]–[10] concentrates on AI generated video quality assessment. Recently, AI-generated 3D content quality assessment [5] has gained significant attention. However, these databases only focus on T23DC. For 3D content generation, three are many work on SI23DC, which can utilize more detail information and generate more satisfying results. In order to address the SI23DCQA, we establish a SI23DCQA database. Table I presents the attributes of existing AIGCQA databases and proposed SI23DCQA database. For the proposed database, the "Prompts" column denotes the number of input single images.

III. DATABASE CONSTRUCTION

A. Image Selection

According to previous AIGCQA work [3], [6], we need to select relatively few images to cover a large number of real user inputs, this is because the generated 3D content will be fine-grained scored. Considering the practical application scenarios of SI23DC, the input single image generally comes from realistic images, AI generated images, and 3D model rendered images, so we collected the above three sets of images, each set contains 100 images with the resolution of 512×512. For realistic images, we collected images of five categories from the Internet, including 30 animals, 20 fruits, 20 objects, 10 people, and 20 vehicles and replaced the background of the images with white; For AI generated images, we asked the ChatGPT to get 20 common animals, household appliances, plants, vehicles and food, and then used the prompt "the front photo of '*object*' with full body, the background area is all white." and DeepFloyd Model to generate 100 images; For 3D model rendered images, we selected 100 non-repetitive 3D models from these datasets [17]–[19] and used MeshLab to render the front view of these 3D models; Fig 1 (a), (b) and (c) show sample images from realistic images, AI generated images or 3D model rendered images respectively.

(a) Realistic Images



(b) AI Generated Images

(c) Model Rendered Images

Fig. 1. Examples of the input images, (a), (b), and (c) are realistic images, AI generated images and 3D model rendered images respectively.

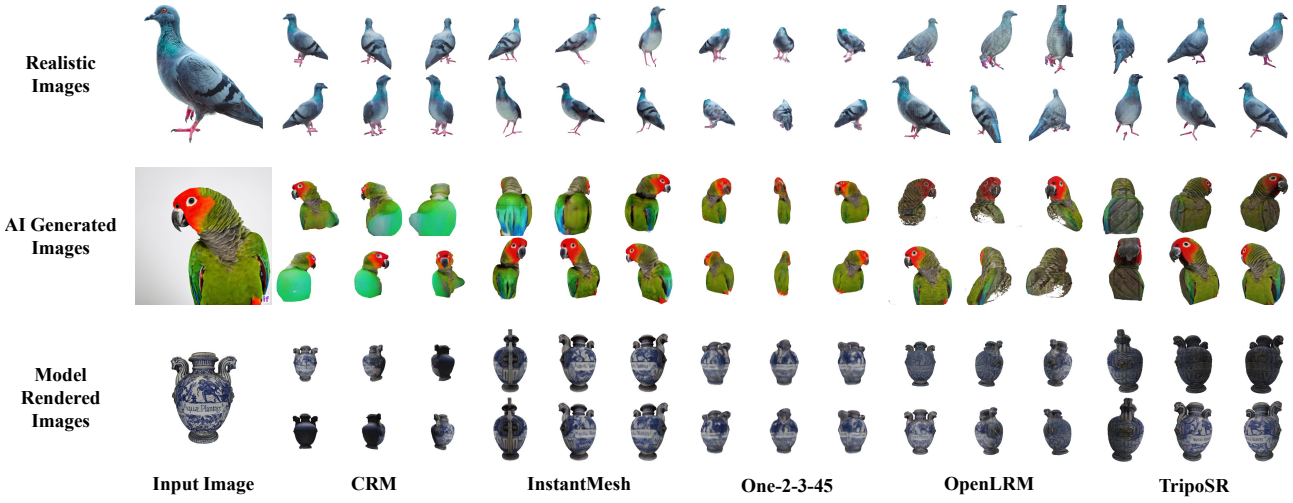


Fig. 2. The samples of SI23DCs from the database. The first to third rows are SI23DC generated from real images, AI generated images, and model rendered images respectively. The leftmost images in each row represent the input images.

B. 3D Content Generation

To ensure the diversity of SI23DCs, our database considered five representative SI23DC generation models. As mentioned above, Current SI23C methods can be divided into two categories, one category utilizes vision-language and 2D diffusion model to predict multi-view images which are used for 3D content reconstruction. We chose One-2-3-45 [11] and CRM [13] as representatives of this category; another category employs an image encoder–3D decoder framework, trained on large-scale 3D datasets [14] to generate 3D content. We selected LRM [12], InstantMesh [15], and TripoSR [16] as representatives of this category. It is worth noting that, One-2-3-45 [11] adopts SDF-based generalizable neural surface for 3D reconstruction, the rest use implicit triplane. For the generation process of SI23DCs, since LRM [12] does not have publicly available code, we chose to use the code of OpenLRM [20], and the remaining methods all used open source code. Additionally, we chose the default configurations and the weight files with the most parameters. Since some methods can not export generated 3D contents as meshes with texture images, we export generated 3D contents as vertex colored meshes, which contains colored vertices $V \in \mathbb{R}^{n\times 6}$ and faces $F \in \mathbb{Z}^{f\times 3}$. These vertex colored meshes are directly used in subsequent subjective experiments. Fig 2 show the samples of SI23DCs from the database.

C. Subjective Experiment

To collect human visual preferences for SI23DCs, we further conducted a well-designed subjective assessment experiment. As highlighted in prior AIGCQA studies [3], [7], the AI generated content are significantly different from human captured or created content, which need to be evaluated from multiple perception perspectives. Hence, in this paper, we propose to evaluate human visual preferences for SI23DCs from three perspectives, including overall, color, and shape. Fig. 3 shows the differences between the selected color and shape dimensions, which further manifests the importance,

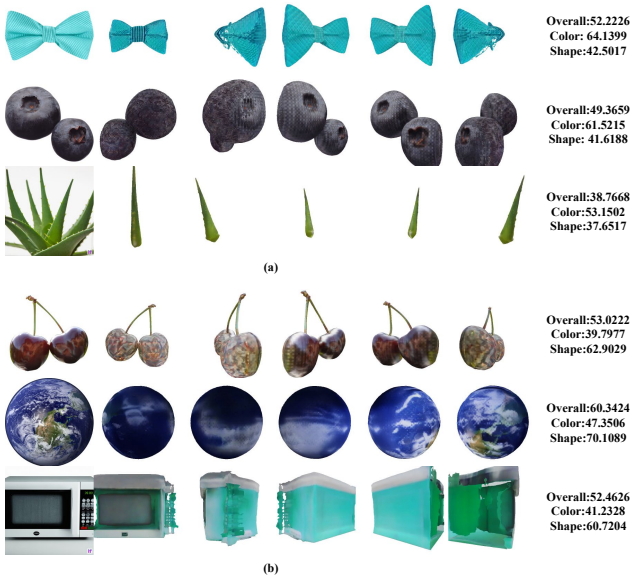


Fig. 3. The illustration of the differences between the two dimensions of color and shape. For each row, the left image is input image and the rest are images rendered from generated 3D content. (a) and (b) show examples of good color correspondence but poor shape correspondence and good shape correspondence but poor color correspondence respectively.

and significance of evaluating SI23DCs from multiple perspectives.

We conducted the subjective experiment following the guidance in ITU-R BT.500-13 [21], which mentions several subjective assessment methodologies, such as single stimulus (SS), double-stimulus impairment scale (DSIS) and paired comparison (PC). Since the input images can be treated as a reference, we adopt DSIS strategy to obtain the subjective quality ratings of generated 3D contents. As shown in Fig 4, for each generated 3D content, subject first views the input image and uses MeshLab software to view the generated 3D content at any angle and distance. There have three sliders ranging from 0 to 5, with a minimum interval of 0.1, are

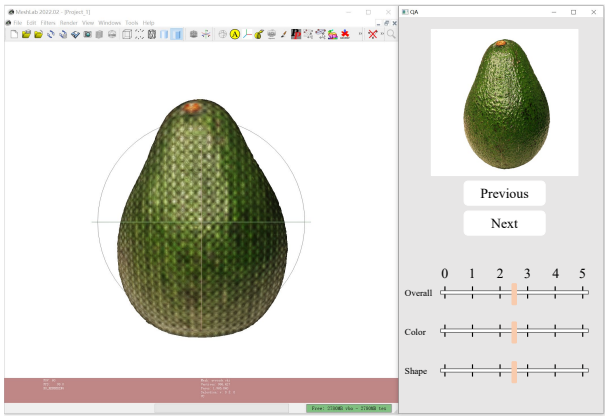


Fig. 4. The subjective quality assessment interface. The left and right part are the SI23DC displayed in MeshLab and the input image respectively.

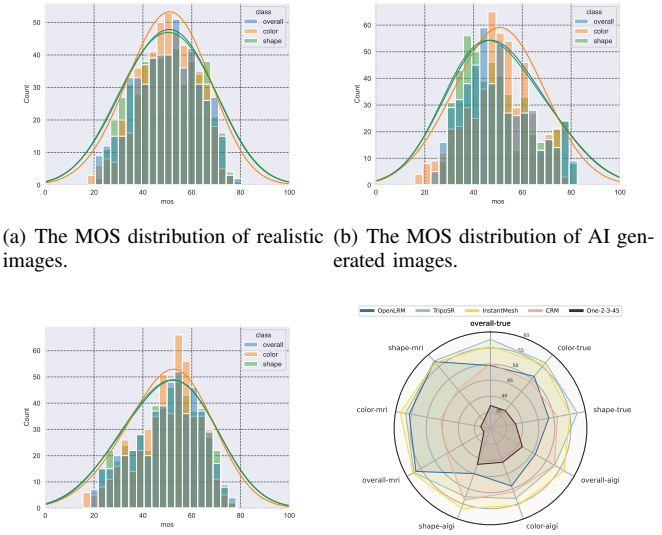
provided for the subject to assign scores for overall, color, and shape. Additionally, two buttons are used to select previous, next generated 3D content. They are seated at a distance of about 2 feet from the monitor, and this viewing distance is roughly maintained during each session. The experimental environment is arranged to simulate a typical indoor home setting with standard lighting conditions. A total of 21 subjects (12 males and 9 females) participated in the subjective experiment, all possessing normal or corrected-to-normal vision. Each participant received detailed experimental instructions prior to engaging in the subjective evaluation. Since the collect input image can be divided in to realistic images, AI generated images and 3D model rendered images, we also divided the subjective experiment into three subsets. For each subsets subjective experiment, subject has 4 hours to label the perception scores of generated 3D contents.

D. Data Processing

Following the data processing used in previous work [22] to conduct the subject rejection and outlier detection. Specifically, for each subset and each evaluation dimension, the kurtosis score of the raw subjective quality ratings for each 3D content is calculated to determine whether it follows a Gaussian or non-Gaussian distribution. For the Gaussian case, the raw score for generated 3D content is regarded as an outlier if it is out side 2 standard deviations (stds) about the mean score of that 3D content; for the non-Gaussian case, it is consider to be an outlier if it outsize $\sqrt{2}0$ stds about the mean score of that 3D content. A subject is removed if more than 3% of his/her evaluations of any dimension are outliers. As a result, only 1 subject is rejected in AI generated images subset. Finally, we convert the raw ratings into Z-scores, which are then linearly scaled to the range $[0, 100]$ and averaged over subjects to obtain the final MOSs as follows:

$$z_{ij} = \frac{m_{ij} - \mu_i}{\sigma_i}, \quad z'_{ij} = \frac{100(z_{ij} + 3)}{6}, \tag{1}$$

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} z'_{ij}, \tag{2}$$



(a) The MOS distribution of realistic images.

(b) The MOS distribution of AI generated images.

(c) The MOS distribution of model rendered images.

(d) The performance of used SI23DC algorithms

Fig. 5. Distributions of the three types MOSs in three subsets and the performance of used SI23DC algorithms.

where m_{ij} is the raw rating given by the i-th subject to j-th generated 3D content, μ_i and σ_i are the mean rating and standard deviation given by subject i respectively, and N is the total number of subjects.

E. Subjective Data Analysis

Based on the results of the subjective experiment, we draw the MOSs distribution for overall, color, and shape dimensions in three subsets and the performance of different generation algorithms, which can be seen in Fig 5. By looking at these figures, some conclusions can be drawn: (1) The overall score of SI23DCs in each subset stays at around 40-60 points, which indicates that the quality of current SI23DCs is not good enough for large-scale application and shows the necessity of SI23DCQA. (2) The MOSs in AI generated image subsets are lower than others, this may because three are some unnatural color distribution and unrealistic structural details in AI generated images. (3) The performance of OpenLRM [20] and InstantMesh [15] surpasses that of other algorithms, while One-2-3-45 [11] performs the worst.

IV. Benchmark Experiment

A. Benchmark Competitors

Considering the traditional 3DQA can be divided into projection-based and model-based methods, we test some NR-IQA methods, including NIQE [23], ILNIQE [24], BRISQUE [25], QAC [26], BPRI [27], BMPRI [28], Resnet [29], Swin Transformer [30], CNNIQA [31], StairIQA [32], HyperIQA [33], NR-VQA methods including MC3-18 [34], R2P1D-18 [34], R3D-18 [34], Swin3D [35], SimpleVQA [36], Fast-VQA [37], DOVER [38], and NR-3DQA methods, including 3D-NSS [39], MM-PCQA [40], GMS-3DQA [41]. These methods contains handcrafted-based, common DNN-based methods.

TABLE II
BENCHMARK PERFORMANCE ON THE PROPOSED SI23DCQA DATABASE. [KEY: **Best**, **Second Best**]

Dimension		Overall			Color			Shape		
Type	**Metric**	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC
NR-IQA	NIQE [23]	0.2183	0.1472	0.2481	0.2230	0.1506	0.2824	0.2103	0.1413	0.2266
	ILNIQE [24]	0.0985	0.0656	0.1796	0.1299	0.0858	0.2178	0.0756	0.0503	0.1560
	BRISQUE [25]	0.2193	0.1444	0.2612	0.2492	0.1642	0.3005	0.1956	0.1285	0.2326
	QAC [26]	0.2383	0.1596	0.2796	0.2486	0.1664	0.3078	0.2274	0.1522	0.2586
	BPRI [27]	0.0493	0.0338	0.1445	0.0554	0.0391	0.1503	0.0548	0.0376	0.1406
	BMPRI [28]	0.1639	0.1088	0.1969	0.1911	0.1276	0.2358	0.1512	0.0997	0.1772
	Resnet-18 [29]	0.6436	0.4622	0.6599	0.6451	0.4641	0.6686	0.6320	0.4517	0.6397
	Resnet-50 [29]	0.5853	0.4204	0.6154	0.5880	0.4208	0.6319	0.5656	0.4049	0.5915
	Swin-T [30]	0.6913	0.5032	0.7068	0.7038	0.5147	0.7241	0.6717	0.4861	0.6822
	Swin-B [30]	0.7151	0.5251	0.7254	0.7219	0.5322	0.7397	**0.6983**	0.5093	**0.7048**
	Swin-L [30]	**0.7222**	**0.5306**	**0.7312**	**0.7371**	**0.5438**	**0.7488**	0.6955	0.5075	0.7011
	CNNIQA [31]	0.4483	0.3086	0.4631	0.4549	0.3137	0.4786	0.4399	0.3025	0.4512
	StairIQA [32]	0.6526	0.4739	0.6680	0.6673	0.4864	0.6838	0.6316	0.4569	0.6462
	HyperIQA [33]	0.6307	0.4520	0.6470	0.6396	0.4594	0.6656	0.6052	0.4317	0.6182
NR-VQA	MC3-18 [34]	0.6498	0.4663	0.6647	0.6490	0.4661	0.6753	0.6350	0.4520	0.6441
	R2P1D-18 [34]	0.6354	0.4556	0.6509	0.6509	0.4659	0.6730	0.6050	0.4288	0.6184
	R3D-18 [34]	0.6365	0.4572	0.6533	0.6524	0.4692	0.6760	0.6271	0.4476	0.6354
	Swin3D-T [35]	0.6844	0.4963	0.6973	0.6882	0.4990	0.7064	0.6676	0.4823	0.6761
	Swin3D-S [35]	0.6999	0.5113	0.7171	0.7025	0.5139	0.7273	0.6856	0.4994	0.6972
	Swin3D-B [35]	0.6430	0.4617	0.6616	0.6418	0.4620	0.6704	0.6273	0.4489	0.6401
	SimpleVQA [36]	0.7576	0.5642	0.7703	0.7799	0.5864	0.7998	0.7294	0.5393	0.7404
	Fast-VQA [37]	0.6987	0.5131	0.7124	0.7326	0.5423	0.7471	0.6958	**0.5098**	0.7006
	DOVER [38]	0.7077	0.5213	0.7177	0.7301	0.5414	0.7472	0.6917	0.5059	0.6953
NR-3DQA	3D-NSS [39]	0.4363	0.3007	0.4463	0.3259	0.2340	0.3887	0.3575	0.2484	0.3952
	MM-PCQA [40]	0.5362	0.3761	0.5520	0.5455	0.3825	0.5724	0.5177	0.3607	0.5298
	GMS-3DQA [41]	0.6399	0.4596	0.6474	0.6579	0.4741	0.6771	0.5999	0.4298	0.6043

B. Experimental Setup

With the help of Pytorch3D [42], we render the generated 3D contents to projection videos, each video contains 120 frames and lasts 4 seconds. For NR-IQA methods, we uniformly sample 12 frames from the video to predict scores and use the average as the final quality score; For NR-VQA methods, we use the rendered videos to predict quality scores; For NR-3DQA methods, we directly use the generated 3D contents to predict quality scores. For the methods that require training, we follow the settings used in previous works [37], [38], we partition the proposed database into training and test sets at a ratio of 4:1. Additionally, we conduct 10 random splits of the database and average the results to ensure unbiased performance comparison. For other methods, we simply operate them on the whole database and report the average performance. Three 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), and Pearson Linear Correlation Coefficient (PLCC). An excellent model should obtain values of SRCC, KRCC, and PLCC close to 1.

C. Performance Discussion

The experimental results are shown in Table II, from which we can make several useful conclusions: (a) For the NR-IQA methods, the DNN-based methods tend to predict quality score more accurate than handcrafted-based methods. This is because the handcrafted-based methods is aimed to assessment the quality of natural image, so they are based on the natural scene statistics (NSS) prior knowledge, which are quite different from the distortions in the rendered images of generated 3D content. (b) SimpleVQA achieves the best performance among all the benchmark competitors in three type MOSs. In our opinions, SimpleVQA utilizes 3D CNN to extract motion features in addition to using 2D CNN for spatial features extraction, which are more relevant to the quality representation of generated 3D content. (c) For NR-3DQA methods, the performance of these methods is more worse than NR-VQA, which may because the features extracted from patch and several projection images are not express the quality of generated 3D content well.

V. CONCLUSION

In this paper, we propose the first SI23DCQA database. We chose 3 types images and 5 common SI23DC algorithms to generate 1500 SI23DCs. Afterward, a subjective study is carried out to collect the subjective quality MOSs for the generated 3D contents. Additionally, Some state-of-the-art NR-IQA, NR-VQA, NR-3DQA methods are chosen for validation on the proposed database. A comprehensive performance discussion is made as well. We hope our work will draw more attention to the quality assessment of generated 3D contents and inspire future research.

REFERENCES

[1] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai, "A perceptual quality assessment exploration for aigc images," 2023.

[2] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin, "Agiqa-3k: An open database for ai-generated image quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[3] Jiarui Wang, Huiyu Duan, Jing Liu, Shi Chen, Xiongkuo Min, and Guangtao Zhai, "Aigciqa2023: A large-scale image quality assessment database for ai generated images: from the perspectives of quality, authenticity and correspondence," in *CICAI*. Springer, 2023, pp. 46–57.

[4] Tengchuan Kou, Xiaohong Liu, Zicheng Zhang, Chunyi Li, Haoning Wu, Xiongkuo Min, Guangtao Zhai, and Ning Liu, "Subjective-aligned dateset and metric for text-to-video quality assessment," *arXiv preprint arXiv:2403.11956*, 2024.

[5] Yingjie Zhou, Zicheng Zhang, Farong Wen, Jun Jia, Yanwei Jiang, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai, "3dgcqa: A quality assessment database for 3d ai-generated contents," *arXiv preprint arXiv:2409.07236*, 2024.

[6] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, et al., "Aigiqa-20k: A large database for ai-generated image quality assessment," *arXiv preprint arXiv:2404.03407*, 2024.

[7] Liu Yang, Huiyu Duan, Long Teng, Yucheng Zhu, Xiaohong Liu, Menghan Hu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet, "Aigcoiqa2024: Perceptual quality assessment of ai generated omnidirectional images," *arXiv preprint arXiv:2404.01024*, 2024.

[8] Iya Chivileva, Philip Lynch, Tomas E Ward, and Alan F Smeaton, "Measuring the quality of text-to-video model outputs: Metrics and dataset," *arXiv preprint arXiv:2309.08009*, 2023.

[9] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan, "Evalcrafter: Benchmarking and evaluating large video generation models," *arXiv preprint arXiv:2310.11440*, 2023.

[10] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou, "Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation," *arXiv preprint arXiv: 2311.01813*, 2023.

[11] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su, "One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan, "Lrm: Large reconstruction model for single image to 3d," *arXiv preprint arXiv:2311.04400*, 2023.

[13] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu, "Crm: Single image to 3d textured mesh with convolutional reconstruction model," *arXiv preprint arXiv:2403.05034*, 2024.

[14] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, et al., "Objaverse-xl: A universe of 10m+ 3d objects," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan, "Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models," *arXiv preprint arXiv:2404.07191*, 2024.

[16] Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao, "Triposr: Fast 3d object reconstruction from a single image," *arXiv preprint arXiv:2403.02151*, 2024.

[17] Yana Nehmé, Johanna Delanoy, Florent Dupont, Jean-Philippe Farrugia, Patrick Le Callet, and Guillaume Lavoué, "Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric," *ACM Transactions on Graphics*, vol. 42, no. 3, pp. 1–20, 2023.

[18] Qi Yang, Joel Jung, Haiqiang Wang, Xiaozhong Xu, and Shan Liu, "Tsmd: A database for static color mesh quality assessment study," in *IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2023, pp. 1–5.

[19] Bingyang Cui, Qi Yang, Kaifa Yang, Yiling Xu, Xiaozhong Xu, and Shan Liu, "Sjtu-tmqa: A quality assessment database for static mesh with texture map," in *ICASSP 2024-2024*. IEEE, 2024, pp. 7875–7879.

[20] Zexin He and Tengfei Wang, "Openlrm: Open-source large reconstruction models," https://github.com/3DTopia/OpenLRM, 2023.

[21] I. T. Union, "Methodology for the subjective assessment of the quality of television pictures," *ITU-R Recommendation BT. 500-11*, 2002.

[22] Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet, "Confusing image quality assessment: Toward better augmented reality experience," *IEEE Transactions on Image Processing (TIP)*, vol. 31, pp. 7206–7221, 2022.

[23] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[24] Lin Zhang, Lei Zhang, and Alan C Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[25] 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.

[26] Wufeng Xue, Lei Zhang, and Xuanqin Mou, "Learning without human scores for blind image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 995–1002.

[27] Xiongkuo Min, Ke Gu, Guangtao Zhai, Jing Liu, Xiaokang Yang, and Chang Wen Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2017.

[28] Xiongkuo Min, Guangtao Zhai, Ke Gu, Yutao Liu, and Xiaokang Yang, "Blind image quality estimation via distortion aggravation," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 508–517, 2018.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.

[31] Le Kang, Peng Ye, Yi Li, and David Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1733–1740.

[32] Wei Sun, Xiongkuo Min, Danyang Tu, Siwei Ma, and Guangtao Zhai, "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training," *IEEE Journal of Selected Topics in Signal Processing*, 2023.

[33] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *CVPR*, June 2020.

[34] 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.

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, "Video swin transformer," in *CVPR*, 2022, pp. 3202–3211.

[36] Wei Sun, Xiongkuo Min, Wei Lu, and Guangtao Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, p. 856–865.

[37] 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," *ECCV*, 2022.

[38] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou Hou, Annan Wang, Wenxiu Sun Sun, Qiong Yan, and Weisi Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *ICCV*, 2023.

[39] 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*, 2022.

[40] 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," *arXiv preprint arXiv:2209.00244*, 2022.

[41] Zicheng Zhang, Wei Sun, Haoning Wu, Yingjie Zhou, Chunyi Li, Zijian Chen, Xiongkuo Min, Guangtao Zhai, and Weisi Lin, "Gms-3dqa: Projection-based grid mini-patch sampling for 3d model quality assessment," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 6, pp. 1–19, 2024.

[42] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari, "Accelerating 3d deep learning with pytorch3d," *arXiv:2007.08501*, 2020.