# PAPS-OVQA: Projection-Aware Patch Sampling for Omnidirectional Video Quality Assessment


1st Chunyi Li
*Shanghai Jiao Tong Univ*
Shanghai, China
lcysyzxdxc@sjtu.edu.cn

2nd Zicheng Zhang
*Shanghai Jiao Tong Univ*
Shanghai, China
zzc1998@sjtu.edu.cn

3rd Haoning Wu
*Nanyang Technological Univ*
Singapore
haoning001@e.ntu.edu.sg

4th Kaiwei Zhang
*Shanghai Jiao Tong Univ*
Shanghai, China
zhangkaiwei@sjtu.edu.cn

5th Lei Bai
*Shanghai AI Laboratory*
Shanghai, China
bailei@pjlab.org.cn

6th Xiaohong Liu
*Shanghai Jiao Tong Univ*
Shanghai, China
xiaohongliu@sjtu.edu.cn

7th Guangtao Zhai
*Shanghai Jiao Tong Univ*
Shanghai, China
zhaiguangtao@sjtu.edu.cn

8th Weisi Lin
*Nanyang Technological Univ*
Singapore
wslin@ntu.edu.sg



***Abstract*—In immersive multimedia systems, the perceptual quality model of omnidirectional video is indispensable. However, to cope with its resolution that is several times higher than ordinary video, the existing omnidirectional video quality assessment (OVQA) models require extremely high computational complexity and usually need to transcode the projection into a certain format. Therefore, to assess the perceptual quality of omnidirectional video effectively, we propose Projection-Aware Patch Sampling (PAPS)-OVQA to process its three common projection formats simultaneously while resizing high-resolution video into patches sampled from uniform grids and finally apply Fragment Attention Network (FANet) to perform quality regression. As a result, we avoid the overhead computational cost of projection transcoding and reduce the complexity of the quality model greatly. Experimental data show that PAPS-OVQA guarantees good performance while retaining high efficiency under different projection formats.**

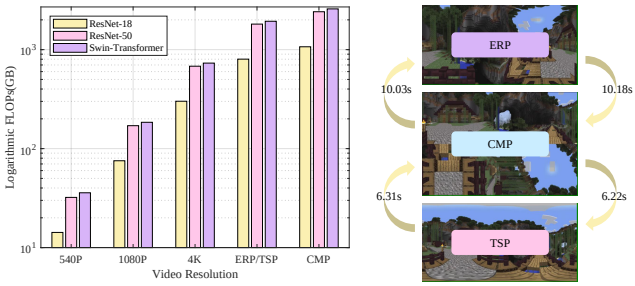


***Index Terms*—omnidirectional video, video quality assessment, patch sampling, 360 video projection**


## I. Introduction

Omnidirectional video [1], [2] is a new type of media content widely used in today's video services. With the help of Head-Mounted (HM) [3] displays, it allows viewers to freely rotate their perspective while watching videos for a more realistic and immersive experience. With the wide application of omnidirectional video in virtual tourism, game development, education, etc., it has become an important part of virtual reality (VR). To cover the user's 360×180° viewing range, the resolution of omnidirectional videos is usually several times that of traditional 2D videos, which places a huge burden on video storage and transmission [4]–[6]. Therefore, there is an urgent need to propose an effective omnidirectional video perceptual quality model to guide video retrieval, thereby optimizing the viewer's experience in the VR world.

However, compared with 2D image/video quality assessment (I/VQA) [9]–[14], designing a convenient and effective Omnidirectional-VQA (OVQA) model is extremely challenging for the following reasons. **Distortion**: The distortion of omnidirectional videos usually comes from splicing and spatial distortion, which is quite different from the distortion mechanism of 2D videos. Directly applying the existing VQA model does not perform satisfactorily on the OVQA task. **Resolution**: The resolution of omnidirectional videos is much higher than 2D videos (usually 8K or 16K), which causes $10^9$ FLOating Points (FLOPs) of operations, resulting in 100× 2D computational costs like Fig.1(a). **Projection**: Omnidirectional video contains multiple projection formats such as Equi-Rectangular Projection (ERP), reshaped Cube-Map Projection (CMP), and Truncated Square Pyramid (TSP). If the model only supports one projection, it will not only be limited in versatility but also cause projection transcoding costs in Fig.1(b) beyond computational costs.




(a) Computational Costs (b) Transcoding Costs

Fig. 1. The OVQA model complexity for different projections. (a)shows the difference in FLOPs required to process 2D or panoramic images under three commonly used networks [7], [8] with batch size 4. (b) lists the transcoding time of three panoramic projections.

## II. Related Work and Contributions

On VQA tasks, from the earliest No Reference (NR) methods such as Brisque [15] and Piqe [16] to advanced methods [17]–[21], they all need to calculate signal fidelity or perform network convolution. On high-resolution OVQA


The work was supported in part by the National Natural Science Foundation of China under Grant 62301310, and in part by the Shanghai Pujiang Program under Grant 22PJ1406800. (Corresponding author: Xiaohong Liu, Guangtao Zhai.)

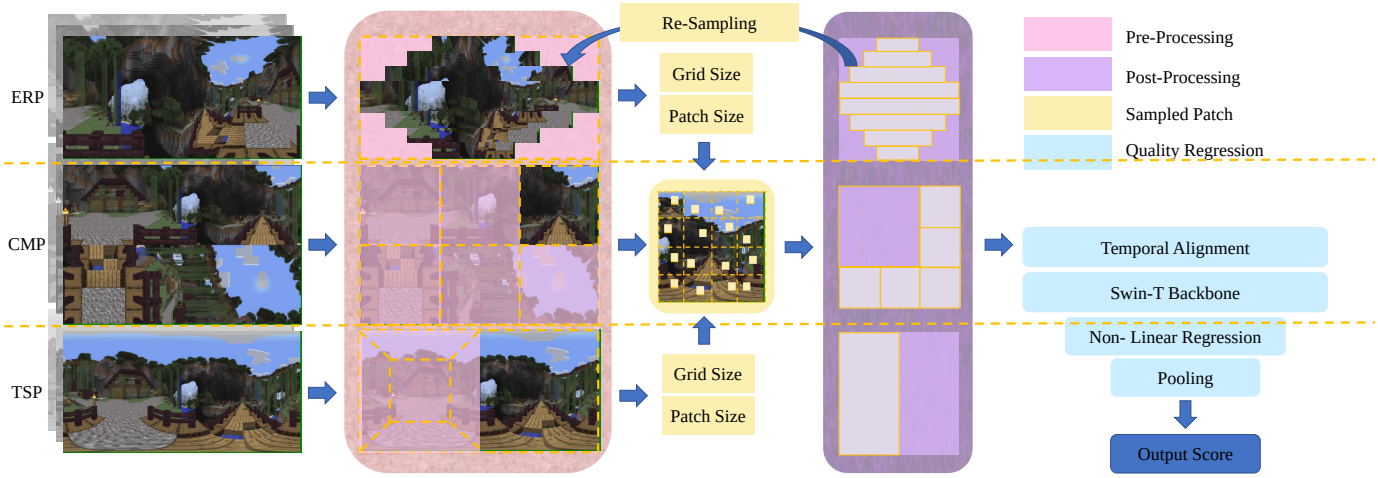



Fig. 2. The framework of the PAPS-OVQA model. Each different projection will undergo different pre-processing, sampling, and post-processing pipelines.

tasks, their time complexity is unacceptable. In recent years, models including FAST-VQA [22] and DOVER [23] apply Patch-Sampling, which greatly reduces computational time by reducing the network input size, but this sampling mechanism performs unsatisfactorily on OVQA tasks. The OVQA models designed for omnidirectional video have greatly improved their performance, but they uniformly transcode the video into CMP [24] or ERP [25] format. Although it reduces the computational time, it brings additional costs of transcoding.

Therefore, to efficiently obtain an accurate perceptual quality [26] under a wide variety of projection formats, we propose the Projection-Aware Patch Sampling (PAPS)-OVQA model with the following contribution:

- We pre-processed each frame collected to ensure that the size of sampled patches is consistent with the size of the spherical projection, thereby capturing distortion information to improve model performance.
- We used patch sampling and performed post-processing after sampling to maximize the utilization of fragments and significantly reduce the computational overhead by reducing the fragment size.
- We use different pipelines for the three projection mechanisms to ensure the model's strong versatility in today's variant VR projection mechanisms.

## III. Proposed Model

### A. Overview

To meet the requirement in Sec. I, our PAPS-OVQA model is mainly divided into two parts: patch embedding and qualify regression. Patch embedding is divided into three submodules: pre-processing, patch-sampling, and post-processing. For videos in ERP, CMP, and TSP projection formats, the sampling grid and patch size depend on each projection's sphere projection area and visual saliency. Patch embedding spatially divides each frame into multiple grids, extracts patches of different sizes in each grid, and then combines them into an entire fragment as network input, thereby reducing the calculation and avoiding the transcoding time. Quality regression uses the Swin-T transformer architecture to extract features and gives the final quality score through nonlinear regression to ensure that performance is not affected by a small input scale.

### B. Patch Embedding

Due to the limited viewing angle of the Human Visual System (HVS), viewers cannot focus on all parts of the omnidirectional video, but on a relatively fixed area [27]. Therefore, the longer an area is viewed, the higher the correlation with subjective quality. To reach a high correlation between objective and subjective quality, pre-processing needs to ensure a consistent actual size of the sampled patch. Meanwhile, the longer an area is viewed, the more likely it is to be sampled.

ERP needs to deal with both aspects. On the one hand, the projected area of the equator on the ERP is much smaller than that of the two poles. A small patch of 16×16 pixels [28] can correspond to a large pole region. On the other hand, the HVS pays far more attention to the content of the equator than to the two poles. At both ends, more patches need to be sampled. Therefore, we consider an ERP projection image, divide it longitudinally into strips $s_{1\sim m}$, and calculate the area corresponding to each strip. Firstly, on a sphere, the area of spherical crown $C_\theta$ with latitude angle $\theta$ is:

$$C_\theta = \int_{\vartheta=\pi/2}^{\theta} 2\pi r \cos\vartheta d\vartheta = 2\pi r^2(1-\sin\theta) \tag{1}$$

where $r$ is the radius of the sphere. Therefore the size of $s_i$ follows:

$$s_i \propto \sin\frac{\pi * i}{m} - \sin\frac{\pi * (i-1)}{m} \tag{2}$$

where $i \in [1, m]$. Based on the proportions of each strips, we can scale the ERP image to a spindle for even sampling. However, to avoid the additional overhead of transcoding the image into a spindle, PAPS-OVQA enlarges each grid and

(a) ERP frame with bitrate (18646k, 3307k, 1216k).

(b) CMP frame with bitrate (33080k, 8678k, 3571k).

Fig. 3. Example of ERP and CMP projection and the sampled patches. Left-to-Right, Up-to-Bottom indicate quality from high to low.

the patches inside and then shrinks them proportionally after splicing. Therefore, the grid and patch size $(g_i, p_i)$ follows:

$$s_i \propto g_i \propto p_i \tag{3}$$

For CMP video, we don't need to justify grid size as the six sides are equal in space. However, the HM data in VQA-ODV shows that the human eye spends more than twice as long viewing the front side as the other sides. Therefore, we need to sample a patch with a larger area in front to achieve the desired effect. Assuming that the perceptual quality of the network's output is linearly related to the size of a certain part of the image, based on the above viewing time ratio, we believe that the length of the front patch $p_{front}$ and other $p_{others}$ parts satisfy:

$$p_{front} = 2p_{others} \tag{4}$$

For TSP video, according to the above CMP analysis, we do not change the grids because the front itself accounts for half of the image. However, considering that the remaining five surfaces occupy a small area, the patch needs to be adjusted. Since the area ratio of the front face to other faces is about 4:1, we also adopt the patch size of (4). After pre-processing, we obtain a patch $P_n$ as the equation below:

$$P_n = I(\frac{x}{g_n}H : \frac{x+p_n}{g_n}H, \frac{y}{g_n}W : \frac{y+p_n}{g_n}W) \tag{5}$$

where $I$ is the original image frame while $(x, y)$ is the coordinate index. $n \in [1, m]$ in ERP, or $\{\text{front}, \text{others}\}$ in CMP and TSP. After sampling each grid, the patch is spliced into the initial fragment, which can be post-processed.

For ERP, because there is no grid in some parts near the two poles, no patch is obtained. To utilize these spaces, we continue patch sampling on the equator, the part of most concern in HVS, until the entire fragment is filled. For CMP, since the front patch is twice as long as the rest, we follow the structure of Fig.2 to let it occupy 4/9 of the area. For TSP, although the front patch is twice the size of other patches, its mapped area on the spherical projection is the same as the rest, so it is reduced by half before slicing. From this, patch embedding can get the complete fragment.

### C. Quality Regression

After avoiding the time-consuming process of projection transcoding, to reduce the time of the quality regression, we use Swin-T Transformer [22], [23] as the backbone extracting the key features from the spatial downsampled patches. In the past VQA tasks Compared with traditional Convolutional Neural Network (CNN) [29], [30], this lightweight structure [31]–[35] has achieved higher performance and lower complexity. For the $t$-th frame of the video, the feature map $F_t$ obtained by Swin-T Transformer is as follows:

$$F_t = \mathcal{T}(P_1 \oplus \cdots \oplus P_n) \tag{6}$$

where Swin-T backbone $\mathcal{T}$ extract the features from the concated $\oplus$ patches $P_1 \cdots P_n$. Thus we report the video's overall quality $Q$ as:

$$Q = \mathcal{P}(\mathcal{NL}(F_1 \oplus \cdots \oplus F_t)) \tag{7}$$

where $\mathcal{P}$ stands for global average pooling, while $\mathcal{NL}$ are several alternating 3D-convolution layers for spatio-temporal feature fusion and a non-linear layer as activation function.

## IV. Expriment

### A. Experiment Setup

The proposed metric is validated on the VQA-ODV [37] and VRVQW [25] databases. VQA-ODV is the most widely used OVQA database for distortion quality assessment, with 540 videos in ERP, CMP, and TSP projection; VR-VQW is a large-scale video quality assessment database, which contains 502 User Generated video samples for aesthetic quality assessment with only ERP projection. Combining the two databases, we can not only verify the effect of PAPS-OVQA on quality evaluation in terms of distortion and aesthetics but also prove its versatility on different projections.

The databases are split randomly in an 80/20 ratio for the training/testing set, and the partitioning and assessment are repeated 10 times for fair comparison and computational complexity, while the average result is reported as the final performance. Our metric is compared with 7 widely-used VQA metric, which shows outstanding performance in previous VQA and OVQA tasks. To avoid excessive computation time, we sample 1 frame/sec as input for IQA metrics.

TABLE I

PERFORMANCE RESULTS ON THE VQA-ODV DATABASE, INCLUDING ERP, CMP, AND TSP VIDEO PROJECTION SUBSETS. THE BEST / SECOND PERFORMANCE RESULTS ARE MARKED IN **RED** / **BLUE**. PROJECTION TRANSCODING-FREE/REQUIRED METHODS ARE NOTED AS ✔/✘.

| Metric | ERP | | | CMP | | | TSP | | | Transcode -Free |
|---|---|---|---|---|---|---|---|---|---|---|
| | SRoCC | KRoCC | PLCC | SRoCC | KRoCC | PLCC | SRoCC | KRoCC | PLCC | |
| Brisque [15] | 0.2952 | 0.2032 | 0.2758 | 0.3151 | 0.2159 | 0.2919 | 0.4201 | 0.3048 | 0.3129 | ✔ |
| Piqe [16] | 0.2716 | 0.1937 | 0.2897 | 0.2116 | 0.1429 | 0.2054 | 0.3810 | 0.2635 | 0.3388 | ✔ |
| V-Bliinds [36] | 0.6947 | 0.4865 | 0.6534 | 0.6890 | 0.4928 | 0.6223 | 0.7187 | 0.5175 | 0.7438 | ✔ |
| FastVQA [22] | 0.5326 | 0.3873 | 0.5238 | 0.5426 | 0.4063 | 0.4802 | 0.5115 | 0.3746 | 0.5380 | ✔ |
| DOVER [23] | 0.5331 | 0.3841 | 0.5674 | 0.5269 | 0.3841 | 0.5432 | 0.4723 | 0.3143 | 0.4466 | ✔ |
| MC360IQA [24] | 0.7483 | 0.5702 | 0.6982 | **0.8241** | **0.6515** | **0.8476** | 0.7308 | 0.5707 | **0.7768** | ✘ |
| ERP-VQA [25] | **0.7823** | **0.6254** | **0.7931** | 0.8033 | 0.6190 | 0.7971 | **0.7514** | **0.5746** | **0.7602** | ✘ |
| **PAPS-OVQA** | **0.8170** | **0.6349** | **0.8154** | **0.8414** | **0.6540** | **0.8078** | **0.7704** | **0.5778** | 0.7326 | ✔ |

TABLE II

APPLYING ONLY THE ERP, CMP, OR TSP PIPELINE OF THE PAPS-OVQA ON THE THREE VQA-ODV VIDEO PROJECTION SUBSETS.

| Pipeline | ERP | | | CMP | | | TSP | | | Time |
|---|---|---|---|---|---|---|---|---|---|---|
| | SRoCC | KRoCC | PLCC | SRoCC | KRoCC | PLCC | SRoCC | KRoCC | PLCC | |
| All | **0.8170** | **0.6349** | **0.8154** | **0.8414** | **0.6540** | **0.8078** | **0.7704** | **0.5778** | **0.7326** | 4.738 |
| ERP | **0.7719** | **0.5968** | **0.7699** | 0.7380 | 0.5746 | 0.7175 | 0.7580 | 0.5778 | 0.7511 | 4.823 |
| CMP | 0.7449 | 0.5524 | 0.7248 | **0.8023** | **0.6032** | **0.7554** | 0.6721 | 0.5016 | 0.6532 | 4.455 |
| TSP | 0.7369 | 0.5651 | 0.7455 | 0.7580 | 0.5714 | 0.7344 | **0.7997** | **0.6095** | **0.7807** | 4.621 |

TABLE III

PERFORMANCE RESULTS ON THE VR-VQW DATABASE AND AVERAGE COMPUTATIONAL TIME ON BOTH DATABASES.

| Metric | VR-VQW | | | Time |
|---|---|---|---|---|
| | SRoCC | KRoCC | PLCC | |
| Brisque [15] | 0.4752 | 0.3288 | 0.4667 | 10.87 |
| Piqe [16] | 0.5087 | 0.3558 | 0.4920 | 16.78 |
| V-Blinds [36] | 0.5031 | 0.3476 | 0.4959 | 372.2 |
| FastVQA [22] | 0.5776 | 0.4081 | 0.5554 | 4.543 |
| DOVER [23] | 0.4489 | 0.3202 | 0.4852 | 4.049 |
| MC360IQA [24] | **0.7703** | 0.5748 | **0.7785** | 10.58 |
| ERP-VQA [25] | 0.7598 | **0.6046** | 0.7772 | 4.125 |
| **PAPS-OVQA** | **0.8324** | **0.6254** | **0.8249** | 4.738 |

We use three common correlation functions, namely Spearman Rank-order Correlation Coefficient (SRoCC), Kendall Rank-order Correlation Coefficient (KRoCC), and Pearson Linear Correlation Coefficient (PLCC), to measure the correlation between metric outputs and the subjective scores. The computational time is verified in seconds on an NVIDIA RTX 4090 GPU.

### *B. Experimental Results and Discussion*

Table I and III show the performance results of PAPS-OVQA and existing methods, where a larger correlation factor indicates a higher consistency with the HVS. For the first three VQA metrics [15], [16], [36], Table III indicates their computational time is all longer than the video itself (10s). Especially for V-bliinds [36] with a computational time of more than 300s, resulting in its inability to conduct the real-time OVQA task. In contrast, the patch-sampling scheme in the method [22], [23] significantly shortens the calculation time. Unfortunately, its sampling is designed for 2D video and ignores the unique distortion of omnidirectional video. Therefore, on all three projections in Table I and III, the SRoCC/PLCC scores are only about 0.5, whose performance is not satisfying. The method [24], [25] designed for omnidirectional video achieves both high performance and low complexity on the OVQA task. Unfortunately, they only support one projection format among ERP/CMP/TSP, and transcoding is required when evaluating videos in other formats. Therefore, their general versatility is limited. Reduced computation time is offset by additional transcoding complexity.

In contrast, our PAPS-OVQA has significant advantages in performance, complexity, and versatility. In the testing sets of Table I and III (12 correlation coefficients in total), we achieved optimal results on 9 of them. In terms of computational complexity, Table III shows that our time consumption is on par with the current fastest methods. Besides, the most noteworthy advantage is transcoding-free, which enables PAPS-OVQA to support three projection formats altogether and outperforms methods that require transcoding.

### *C. Ablation Study*

To validate the contributions of all components in PAPS-OVQA, we also conduct an ablation study and its results are shown in Table II. We use the same pipeline to process ERP/CMP/TSP videos unitedly, instead of case-by-case. The results show that each pipeline can only handle one kind of projection, and only the combination of three pipelines can improve the performance of all projections.

## V. CONCLUSION

In this study, we target the challenge of measuring the perceptual quality of omnidirectional video with different projections, and PAPS-OVQA, a projection transcoding-free metric is proposed. Experiments show that PAPS-OVQA achieves the best results in two databases across three correlation measures with a low time complexity, which can optimize the utilization of immersive multimedia services.

## REFERENCES


[1] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yucheng Zhu, Yi Fang, and Xiaokang Yang, "Perceptual quality assessment of omnidirectional images," in *IEEE international symposium on circuits and systems (ISCAS)*, 2018.

[2] Huiyu Duan, Guangtao Zhai, Xiaokang Yang, Duo Li, and Wenhan Zhu, "Ivqad 2017: An immersive video quality assessment database," in *International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2017.

[3] Li Yang, Mai Xu, Shengxi Li, Yichen Guo, and Zulin Wang, "Blind vqa on 360° video via progressively learning from pixels, frames, and video," *IEEE Transactions on Image Processing*, 2022.

[4] Chunyi Li, Haoyang Li, Ning Yang, and Dazhi He, "A pbch reception algorithm in 5g broadcasting," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2022.

[5] Haoyang Li, Dazhi He, Chunyi Li, Runnan Liu, Yin Xu, Yihang Huang, and Yunfeng Guan, "Aliasing-elimination channel estimation for cas reception," *IEEE Transactions on Broadcasting*, 2021.

[6] Haoyang Li, Dazhi He, Chunyi Li, Runnan Liu, Yin Xu, Yihang Huang, and Yunfeng Guan, "On the aliasing-elimination for cas channel estimation," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2021.

[7] 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.

[8] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[9] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai, "A perceptual quality assessment exploration for aigc images," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2023.

[10] 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*, 2023.

[11] Yixuan Gao, Yuqin Cao, Tengchuan Kou, Wei Sun, Yunlong Dong, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai, "Vdpve: Vqa dataset for perceptual video enhancement," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2023.

[12] Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[13] Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang, "Exploring the naturalness of ai-generated images," arXiv preprint arXiv:2312.05476, 2023.

[14] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Topiq: A top-down approach from semantics to distortions for image quality assessment," arXiv preprint arXiv:2308.03060, 2023.

[15] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on image processing*, 2012.

[16] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S Channappayya, and Swarup S Medasani, "Blind image quality evaluation using perception based features," in *twenty first national conference on communications (NCC)*, 2015.

[17] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, et al., "Q-bench: A benchmark for general-purpose foundation models on low-level vision," arXiv preprint arXiv:2309.14181, 2023.

[18] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al., "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," arXiv preprint arXiv:2311.06783, 2023.

[19] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al., "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," arXiv preprint arXiv:2312.17090, 2023.

[20] Zicheng Zhang, Haoning Wu, Zhongpeng Ji, Chunyi Li, Erli Zhang, Wei Sun, Xiaohong Liu, Xiongkuo Min, Fengyu Sun, Shangling Jui, et al., "Q-boost: On visual quality assessment ability of low-level multi-modality foundation models," arXiv preprint arXiv:2312.15300, 2023.

[21] Chunyi Li, Haoning Wu, Zicheng Zhang, Hongkun Hao, Kaiwei Zhang, Lei Bai, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and Guangtao Zhai, "Q-refine: A perceptual quality refiner for ai-generated image," arXiv preprint arXiv:2401.01117, 2024.

[22] 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 *European Conference on Computer Vision*, 2022.

[23] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[24] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Huiyu Duan, and Siwei Ma, "Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[25] Wen Wen, Mu Li, Yiru Yao, Xiangjie Sui, Yabin Zhang, Long Lan, Yuming Fang, and Kede Ma, "Perceptual quality assessment of virtual reality videos in the wild," arXiv preprint arXiv:2206.08751, 2022.

[26] Guangtao Zhai and Xiongkuo Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, 2020.

[27] Xinhui Huang, Chunyi Li, Abdelhak Bentaleb, Roger Zimmermann, and Guangtao Zhai, "Xgc-vqa: A unified video quality assessment model for user, professionally, and occupationally-generated content," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2023.

[28] Haoning Wu, Chaofeng Chen, Liang Liao, Jingwen Hou, Wenxiu Sun, Qiong Yan, Jinwei Gu, and Weisi Lin, "Ieee transactions on pattern analysis and machine intelligence," *IEEE TPAMI*, 2023.

[29] Chunyi Li, May Lim, Abdelhak Bentaleb, and Roger Zimmermann, "A real-time blind quality-of-experience assessment metric for http adaptive streaming," in *IEEE International Conference on Multimedia and Expo*, 2023.

[30] Chunyi Li, Zicheng Zhang, Wei Sun, Xiongkuo Min, and Guangtao Zhai, "A full-reference quality assessment metric for cartoon images," in *IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, 2022.

[31] 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.

[32] Zicheng Zhang, Wei Sun, Yingjie Zhou, Haoning Wu, Chunyi Li, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin, "Advancing zero-shot digital human quality assessment through text-prompted evaluation," arXiv preprint arXiv:2307.02808, 2023.

[33] Zicheng Zhang, Yingjie Zhou, Long Teng, Wei Sun, Chunyi Li, Xiongkuo Min, Xiao-Ping Zhang, and Guangtao Zhai, "Quality-of-experience evaluation for digital twins in 6g network environments," *IEEE Transactions on Broadcasting*, 2024.

[34] Tengchuan Kou, Xiaohong Liu, Wei Sun, Jun Jia, Xiongkuo Min, Guangtao Zhai, and Ning Liu, "Stablevqa: A deep no-reference quality assessment model for video stability," arXiv preprint arXiv:2308.04904, 2023.

[35] Yunlong Dong, Xiaohong Liu, Yixuan Gao, Xunchu Zhou, Tao Tan, and Guangtao Zhai, "Light-vqa: A multi-dimensional quality assessment model for low-light video enhancement," arXiv preprint arXiv:2305.09512, 2023.

[36] Michele A Saad, Alan C Bovik, and Christophe Charrier, "Blind prediction of natural video quality," *IEEE Transactions on image Processing*, 2014.

[37] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang, "Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018.