Perceptual Quality Assessment of Omnidirectional Images

Yuming Fang[1], Liping Huang[1], Jiebin Yan[1*], Xuelin Liu[1], Yang Liu[2]

[1] Jiangxi University of Finance and Economics, Nanchang, China
[2] SANY Heavy Industry Co., Ltd., China
fa0001ng@e.ntu.edu.sg, {lphuang19, jiebinyan, xuelinliu-bill}@foxmail.com, liuy2655@sany.com.cn

Abstract

Omnidirectional images, also called 360° images, have attracted extensive attention in recent years, due to the rapid development of virtual reality (VR) technologies. During omnidirectional image processing including capture, transmission, consumption, and so on, measuring the perceptual quality of omnidirectional images is highly desired, since it plays a great role in guaranteeing the immersive quality of experience (IQoE). In this paper, we conduct a comprehensive study on perceptual quality of omnidirectional images from both subjective and objective perspectives. Specifically, we construct the largest so far subjective omnidirectional image quality database, where we consider several key influential elements, *i.e.*, realistic non-uniform distortion, viewing condition, and viewing behavior, from the user view. In addition to subjective quality scores, we also record head and eye movement data. Besides, we make the first attempt by using the proposed database to train a convolutional neural network (CNN) for blind omnidirectional image quality assessment. To be consistent with the human viewing behavior in the VR device, we extract viewports from each omnidirectional image and incorporate the user viewing conditions naturally in the proposed model. The proposed model is composed of two parts, including a multi-scale CNN-based feature extraction module and a perceptual quality prediction module. The feature extraction module is used to incorporate the multi-scale features, and the perceptual quality prediction module is designed to regress them to perceived quality scores. The experimental results on our database verify that the proposed model achieves the competing performance compared with the state-of-the-art methods.

Introduction

With the advancement of 5G commercialization, virtual reality (VR) has been used in many applications with new vitality, and omnidirectional images have drawn much attention. Omnidirectional images are taken by cameras with multiple lenses covering the entire 360° scene. Multiple images captured simultaneously by different lenses are then stitched together to form an omnidirectional image. An omnidirectional image is shown in spherical form covering 360×180 viewing range, and thus we can freely explore the scene with the help of head-mounted displays (HMDs). During the exploration, we can get the quality of experience as if we are in the real world. However, this immersive experience puts high demands on the fidelity and resolution of omnidirectional images, bringing a huge burden on storage and transmission. In each stage of omnidirectional image processing (Wang and Rehman 2017), distortions may be introduced due to the diverse environment conditions, algorithm defects, *etc.*, leading to degrade users' Quality of Experience (QoE). How to measure the quality of omnidirectional images accurately plays a critical role for the system optimization.

The research on omnidirectional image quality assessment (OIQA) can be divided into two categories: subjective and objective OIQA. Subjective OIQA refers to constructing large-scale databases as well as collecting human ratings. There have been several omnidirectional image quality (OIQ) databases proposed, *e.g.* (Upenik, Řeřábek, and Ebrahimi 2016), (Sun et al. 2017), and (Duan et al. 2018). Most of the existing OIQA databases focus on the visual quality of omnidirectional images with uniformly distributed distortions, while ignoring the impact of user viewing behavior. Existing objective OIQA methods can be classified into three categories: sampling-related (Yu, Lakshman, and Girod 2015; Sun, Lu, and Yu 2017; Zakharchenko, Choi, and Park 2016), patch-based (Kim, Lim, and Ro 2020; Lim, Kim, and Ra 2018) and viewport-based (Sun et al. 2020; Xu, Zhou, and Chen 2021). Sampling-related methods are mainly designed based on existing 2D quality metrics such as peak signal to noise (PSNR) and structural similarity (SSIM), which are feasible with available reference information. Patch-based methods ignore the fact that subjects explore an omnidirectional image from viewports rather than patches. Although viewport-based methods can well solve the aforementioned issue, most of the viewport-based methods do not regard the exploration of omnidirectional images as a continuous process, which may lead to unsatisfied results. (Sui et al. 2021) made an initial attempt to investigate the relationship between user viewing behavior and the perceived quality of omnidirectional images. They built an OIQ database and designed an OIQA method by taking full advantage of the temporal properties. However, the scale of this dataset is small, and the omnidirectional images are only with two types of distortions, while does not

*Corresponding author.

Dataset	Year	No.Ref	No.Dist	Resolution	Format	Distortion Type
OIQA (Duan et al. 2018)	2018	16	320	11332×5666,13320×6660	PNG, JPG	JPEG, JP2K, GB, GN
CVIQD (Sun et al. 2017)	2018	16	528	4096×2048	PNG	JPEG, H.264/AVC, H.265/HEVC
MVAQD (Jiang et al. 2021)	2021	15	300	4K, 5K, 6K, 7K, 8K, 10K, 12K	-	JPEG, JP2K, HEVC, WN, GB
IQA-ODI (Yang et al. 2021)	2021	120	960	7680×3840	JPG	JPEG, Projection
SOLID (Xu et al. 2018)	2018	6	312	8192×4096	PNG	JPEG, BPG
LIVE 3D IQA (Chen et al. 2020)	2019	15	450	4096×2048	PNG	GB, GN, Downsampling, VP9 compression, H.265
Ours	2021	258	1032	8192×4096	PNG	GB, GN, BD, ST

Table 1: Summary of the existing omnidirectional image quality assessment databases.

take into account the non-uniform distortions caused by non-synchronization between different lenses of the camera. The objective method will be limited when the user's browsing data is not available. To tackle the aforementioned issues, in this paper, we establish a large-scale OIQ database with diverse content and propose a convolutional neural network (CNN) based blind OIQA method by considering viewing behaviors and viewing conditions.

Our contributions include:

- A large-scale OIQ database with diverse content is built. The proposed OIQ database consists of 258 reference images with rich scenes, and associated 1032 non-uniform distorted images which are generated by considering realistic factors.
- An extensive subjective experiment is conducted. We collect both subjective ratings and human viewing behavior data under different conditions. Specifically, rigorous data processing operations are performed to ensure the authenticity of the proposed database, and the relationship between viewing conditions, non-uniform distortions and subjective quality scores are analyzed to further understand the process by which humans perceive the quality of omnidirectional images.
- A novel CNN-based blind OIQA model is proposed. Both viewing behaviors and viewing conditions are taken into account to get a more reasonable result. Through comparing with the state-of-the-arts, we demonstrate the superiority of the proposed model.

Related Work

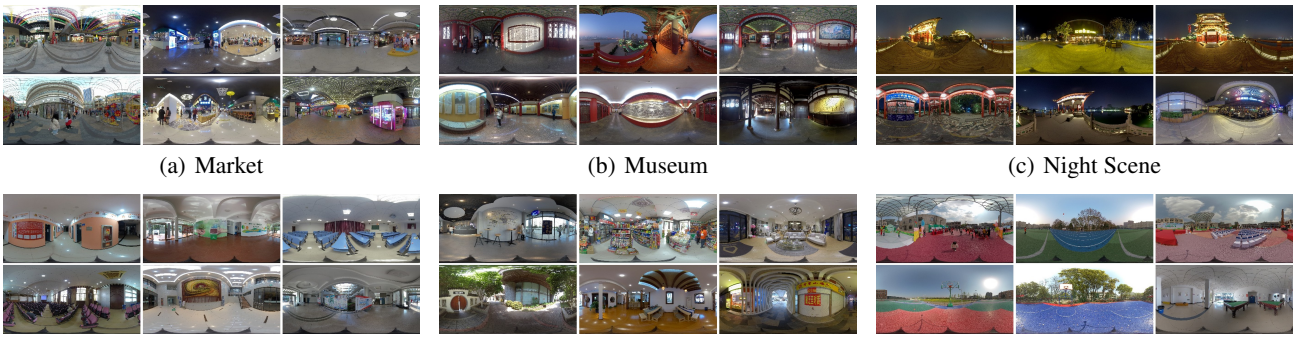
In this section, we first introduce current public subjective omnidirectional image quality (OIQ) databases, and then we describe objective OIQA models in detail.

Subjective OIQ Databases

To facilitate the development of OIQA, constructing OIQ database has always been the previous step before designing objective OIQA models. A subjective OIQ database can be used as: 1) a platform for physiological experiments, whose findings are often used as the guidance for designing objective OIQA models; 2) a benchmark for testing objective OIQA models. (Upenik, Řeřábek, and Ebrahimi 2016) built an OIQ database to study the effect of compression and projection on the visual quality of omnidirectional images. (Sun et al. 2017) constructed an OIQ database with a focus on the influence of coding scheme. This database contains 16 original images and 528 distorted images, and 20 subjects are invited in subjective experiments. (Duan et al. 2018) proposed an OIQ database with more comprehensive information. Specifically, this database contains 16 original omnidirectional images and 320 distorted omnidirectional images with associated subjective ratings. Besides, this database provides subjects' head movement and eye movement data. (Yang et al. 2021) constructed an OIQ database, which mainly considers the impact of JPEG compression and mapping format on OIQ. This database contains 120 high-quality reference omnidirectional images, which are used to generate 960 distorted omnidirectional images with four mapping formats. Considering the re-projected distortion changes between equirectangular projection (ERP) image and spherical projection image, (Jiang et al. 2021) built an OIQ database with multi-distortion. In addition to commonly used general 2D omnidirectional images, researchers have also studied visual quality of 3D omnidirectional images. 3D omnidirectional image differs from general 2D omnidirectional image in that it consists of two views (left- and right-views). (Xu et al. 2018) proposed a 3D OIQ database, which contains 6 high-quality 8K stereoscopic omnidirectional images and 156 distorted versions with different levels of depth and BPG compression distortion. (Chen et al. 2020) proposed a 3D OIQ database, which contains 16 pristine 3D omnidirectional images and 450 distorted 3D omnidirectional images. The distortion types include Gaussian noise, Gaussian blur, downsampling, stitching, VP9 compression, and H.265 compression. Table 1 summaries and compares existing databases for omnidirectional images.

Objective OIQA Models

In the past few years, many OIQA methods, such as S-PSNR (Yu, Lakshman, and Girod 2015), WS-PSNR (Sun, Lu, and Yu 2017), and CPP-PSNR (Zakharchenko, Choi, and Park 2016), was proposed by combining 2D quality evaluation methods (SSIM (Wang et al. 2004) or PSNR) with the characteristics of omnidirectional images. However, PSNR related method does not fully accord with the actual perception of the human visual system (HVS). Some researchers further choose SSIM which is more suitable for human eye perception to evaluate omnidirectional image quality, such as S-SSIM (Chen et al. 2018) and WS-SSIM (Zhou et al. 2018). Both S-PSNR and S-SSIM methods calculate the local quality scores of omnidirectional images based on spherical projection. WS-PSNR and WS-



(a) Market (b) Museum (c) Night Scene

(d) Academic Building (e) Shop (f) Activity Place

Figure 1: Sample images in our database.

SSIM try to reduce the influence of distortion introduced by the mapping transformation from sphere to the 2D plane on the calculation of quality score by introducing the weighting factor. CPP-PSNR projects the omnidirectional image to the Craster parabolic projection plane and then calculates the PSNR. This method ensures the uniformity of sampling density, but the computational cost is relatively high.

(Kim, Lim, and Ro 2020) proposed an OIQA method which consists of two parts, including a predictor module and a guider module. The guider module is trained to distinguish the predicted score and ground truth accurately, and the predictor module predicts the score under the supervision of the guider. Considering that users explore omnidirectional images through viewports rather than patches, many viewport-based approaches have been proposed for OIQA. (Zhou et al. 2021a) designed a no-reference OIQA model which uses multi-frequency information as well as local and global features to measure quality degradation of omnidirectional images. (Sun et al. 2020) proposed a viewport-based multi-channel CNN-based model. This method maps the omnidirectional image into six cubemap faces of equal size. Then, six parallel networks are designed for feature extraction, which is further used for quality prediction of the omnidirectional image. (Zhou et al. 2021b) proposed a distortion discrimination assisted dual-stream network for omnidirectional image quality assessment. Different from the above methods, some researchers considered the interaction between different viewports when people browse 360-degree scenes. (Xu, Zhou, and Chen 2021) assumed that people would generate local and global impressions when observing 360-degree images. Thus, they proposed to use graph convolutional networks to obtain the relationship between different viewports and incorporated this into the local perception feature extraction module. Besides, a top-performing 2D IQA approach (Zhang et al. 2020) is used to extract global features from ERP images.

Benchmark

Database Construction

We collect 258 reference images with diverse content using an Insta360 Pro2 camera. As shown in Figure 2, the proposed database contains 16 different scenes, including

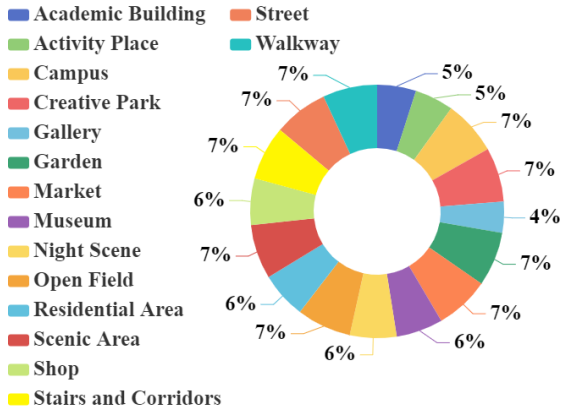
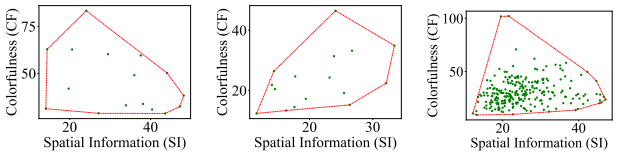


Figure 2: Percentage of various scenarios in the proposed omnidirecitonal image database.



(a) OIQA Database (b) CVIQD Database (c) Our Database

Figure 3: The scatter diagram of Spatial Information (SI) and Colorfulness (CF) of three omnidirectional image databases.

night scene, market, museum, *etc*. Some visual examples are shown in Figure 1. To give a more intuitive comparison on diversity against the existing OIQ databases, we also provide the scatter plots of spatial information and color information, which are quantified by SI (Spatial Information) (ITU-T RECOMMENDATION 1999) and CF (Colorfulness) indexes (Hasler and Suesstrunk 2003) respectively. From Figure 3, we can clearly observe that the proposed OIQ database is more diverse than other existing ones. There are many aspects involved in the overall QoE of omnidirectional images due to the additional immersive content. The immersive experience can be achieved by using the head-mounted

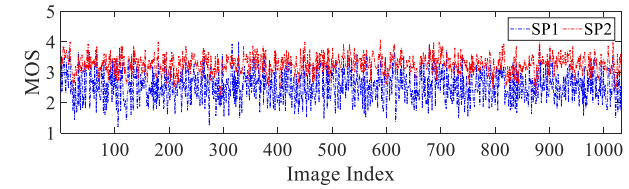
display (HMD) to access the viewport containing the content of interest through head movement. The distorted local regions (with non-uniformed distortion) in the omnidirectional image are much eye-catching, and even affect the perceptual quality of the whole image, since the user concentrates on the viewport (a small part of the omnidirectional image). The goal of our subjective experiment is to study how non-uniformed distortion, viewing behavior and viewing condition affect the perceived quality of omnidirectional images. Therefore, four types of non-uniform distortion are considered in our database, including Gaussian blur, Gaussian noise, brightness discontinuity and stitching, with 3 levels of each type. Gaussian blur and Gaussian noise are used to simulate the distortion produced by the camera during shooting. Since they are usually generated by the camera sensor, we add these two kinds of noise to an image of a lens that contains a salient object separately, and then stitch them back together to form an omnidirectional image. Brightness discontinuity usually manifests as a large difference in brightness between adjacent camera lenses. We simulate it by adding three different degrees of brightness distortion to the image of a lens and then stitching the multi-lens images into a whole. Stitching distortion usually occurs when the stitching algorithm does not successfully handle the inconsistencies among the cameras in the multicamera rig. In our database, we use Nuke (a popular stitching software) to generate different levels of stitching distortion. More specifically, six fisheye images are imported into Nuke to generate a stitched image, and then we adjust the distortion parameters of a camera lens to obtain the distorted images.

Conditions

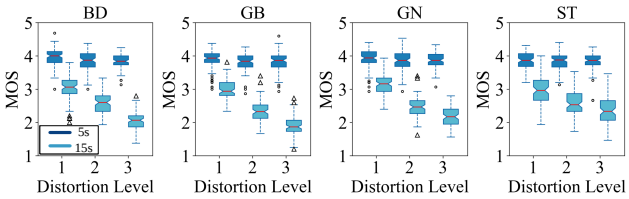
To analyze the effect of the starting point on the perception of visual quality, for each scene, we set two viewing starting points: a good starting point and a bad starting point. The bad starting point means that each user starts viewing the omnidirectional image from the distorted area, while the good starting point means that each user starts viewing from the high-quality area (located at the opposite of the distortion area). In order to investigate the changes of viewing behavior during the exploration of scenes, the viewing process of a single scene is further divided into two phases: five seconds and fifteen seconds. Totally, there are four viewing conditions.

Procedure

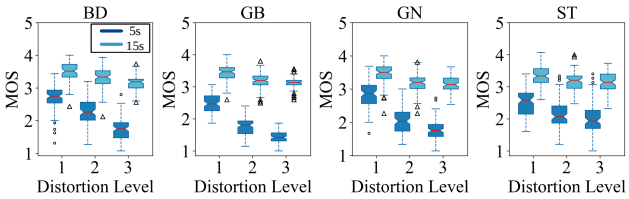
Subjective experiments are performed based on the single stimulus continuous quality evaluation method (BT 2002). The equipment used to show the subjects of the omnidirectional image content includes an HTC VIVE VR headset for tracking the users' head movement data and a high-performance computer to support the operation of the Unity game engine. We invite more than 120 subjects aged from 18 to 26 to participate in our subjective experiments, where each subject is asked to sit in a swivel chair to observe the scenes. Before starting the experiments, subjects can freely calibrate the HMD and are verbally informed about the procedure of the experiment. After that, a training phase is conducted to familiarize subjects with distortion types and



(a) Overall score of images for starting point I and II



(b) Starting point I



(c) Starting point II

Figure 4: MOSs under different viewing conditions.

the scoring procedure. Subjects are divided into two groups and rate the quality of the omnidirectional image from two different starting point sequences. Notably, each image is viewed for 15 seconds, and when the subject watches for five seconds, a tone in the HMD prompts the subject to give a judgment on the quality of this period. After fifteen seconds, the subject has to give the overall perceived quality of the image. Images are rated on a continuous rating scale in [1, 5], where 1 represents the worst quality.

Subjective Data Analysis

Figure 4 (a) shows the overall score of each image after viewing for 15 seconds from starting point I (*i.e.*, the good starting point) and starting point II (*i.e.*, the bad starting point) respectively. From this figure, we can find a significant difference between the overall scores obtained from the same image when viewed from different starting points. This indicates that the starting point has an effect on the perception of image quality. A combination of multiple factors often determines differences in MOSs, thus we employ ANOVA to further analyse the effect of multiple factors on the overall score of the images (viewed for 15 seconds). As can be seen from the Table 2, the p value obtained from ANOVA analysis for distortion type, starting point and scene are statistically significant (p<0.05), which means these factors significantly affect the overall visual scores. Additionally, the interaction of both distortion type and starting point also affects the perception of image quality. Sub-figures (b) and (c) in Figure 4 show the boxplots of MOSs under different distortion types and different viewing conditions. As

$Source$	SS	$d.f.$	MS	F	p
Distortion Type	3.622	3	1.207	6.93	≈ 0
Starting Point	253.6	1	253.6	1456.29	≈ 0
Scene	10.012	15	0.667	3.83	≈ 0
Distortion Type$\times$ Starting Point	4.961	3	1.654	9.5	≈ 0
Distortion Type$\times$ Scene	10.174	45	0.226	1.3	0.0896
Starting Point$\times$ Scene	1.176	15	0.078	0.45	0.9638
Distortion Type$\times$ Starting Point$\times$ Scene	3.734	45	0.083	0.48	0.9988
Error	337.137	1936	0.174		
Total	632.695	2063			

Table 2: ANOVA results for MOSs for the viewing time of 15 seconds. SS: sum of squares. MS: mean square. $d.f.$: degree of freedom. F: F value. p: p value.

can be seen from the sub-figures, the median line in three dark blue boxes (corresponding to three different levels of distortion) of various distortion types is very close to each other when viewed for 5 seconds from starting point I, while the light blue ones vary with the level of distortion. When viewed from starting point II, the median line of the dark blue boxes in each distortion types are more variable. It indicates that the distortion levels have much more impact on subjective quality of 5 seconds in each distortion types of starting point II.

The Proposed OIQA Model

Based on the proposed database, we make the first attempt to train a CNN for blind OIQA. The framework of the proposed model is illustrated in Figure 5. Specifically, several viewports extracted from an omnidirectional image are adopted as inputs of the proposed model. A multi-scale CNNs-based feature extraction module and a perceptual quality prediction module in the proposed model are used to extract the multi-scale features and regress them to the perceived quality score, respectively.

Viewport Extraction

When humans view an omnidirectional image using a VR device, the equirectangular image is transformed into a 3D sphere with spherical coordinates, and the visual content is then rendered as a viewport which is determined by the viewing angle and the field of view (FOV) of the VR device. With the human's head moving, the visible content will be changed. Inspired by human viewing behaviors, we opt to extract viewport images to assess the perceptual quality of the omnidirectional images. Rectilinear projection is adopted to generate viewports and the detailed procedure can be found in (Ye, Alshina, and Boyce 2017). Consisted with the FOV provided in HTC VIVE, we set the FOV as 110°. As suggested in (Xu et al. 2019), we obtain 20 viewports for an omnidirectional image by extracting viewports uniformly distributed over the sphere. In order to incorporate the user viewing conditions naturally in the proposed model, a starting point ϕ and an exploration time T for an omnidirectional image are also used. In brief, the training data is composed of $D = \left\{OI_n, VPI_n^1, \cdots, VPI_n^M, \phi_n, T_n\right\}_{n=1}^N$, where M and N denote the number of the extracted viewports of an omnidirectional image and the training data, respectively. M is set to 20.

Model Construction

CNNs have achieved success in many research fields, such as image recognition (He et al. 2016), detection (Fang et al. 2019) and segmentation (Yan et al. 2021), and some CNNs based models have been proposed for different visual tasks in recent years. Due to the excellent generalization ability of these models, they are applied to extract visual features for image quality assessment (Ding et al. 2020; Zhang et al. 2021; Su et al. 2020; Sun et al. 2020). We choose a variant of ResNet-50 (He et al. 2016) as the backbone for constructing the multi-scale feature extraction network S, which is used to extract multi-scale content features in (Su et al. 2020). Given an omnidirectional image OI and the corresponding viewport images $\{VPI^m\}_{m=1}^M$, we model the procedure of multi-scale feature extraction as follows:

$$f^m = S(VPI^m, \theta_s) \tag{1}$$

where θ_s denotes the network parameters of the multi-scale feature extraction module. The multi-scale content features include local and global features. More detail of semantic feature extraction network structure can be found in (Su et al. 2020). However, different from (Su et al. 2020), we only use the semantic feature extraction network for feature extraction and remove the subsequent network by taking into account the characteristics of omnidirectional images. The extracted multi-scale features are then concatenated with the user viewing conditions, which will further instruct the training of perceptual quality prediction network R. The predicted scores of all viewports in an omnidirectional image are generated by a simple quality prediction network consisting of two fully connected layers. We model the task of quality prediction as follows:

$$q^m = R(concat(f^m, \phi, T), \theta_r) \tag{2}$$

where θ_r represents the network parameters of the perceptual quality prediction module. Finally, the overall perceptual quality of an omnidirectional image can be calculated by averaging the predicted scores of all viewports, which can be described as:

$$q = \frac{1}{M}\sum_{m=1}^{M} q^m \tag{3}$$

According to user's viewing conditions and behaviors in the VR device (Fang, Zhang, and Lmamoglu 2018; Sui et al. 2021), the importance of each viewport image extracted in an omnidirectional image is different for the final perceived quality score. Therefore, some pooling policies can be adopted to assign weights for different viewport images in the image quality regression (Sui et al. 2021). Here, as

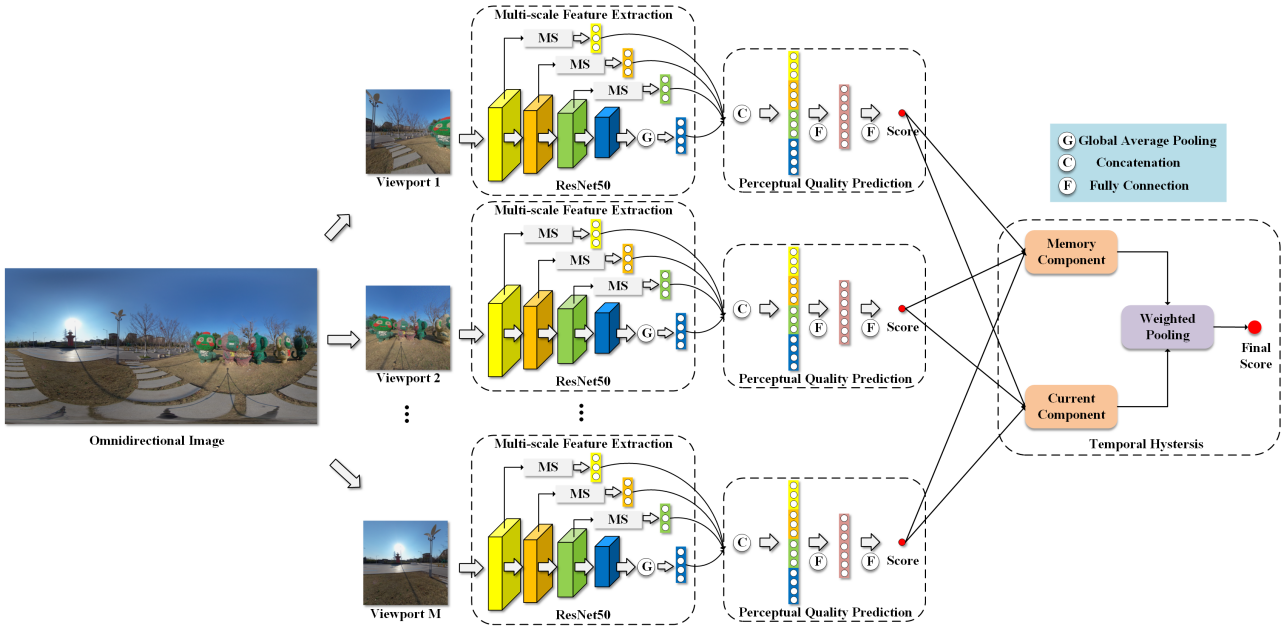


Figure 5: The framework of proposed method. It consist of two main modules. The multi-scale feature extraction module is used to extract multi-scale content features from viewports, and the perceptual quality prediction module regress the concatenated features to the perceived quality scores. The final omnidirectional image quality score is obtained by using simple average pooling strategy. Specifically, MS represents multi-scale subblock that is composed of a series of operations.

shown in Eq. 3, we employ the averaging pooling as the default pooling strategy. Note that other advanced temporal pooling strategies (Li, Jiang, and Ming 2019; Fang et al. 2021; Yan et al. 2021) can be also adopted here.

We minimize L_1 loss for network optimization, and the loss function over the training set is defined as follows:

$$L = \frac{1}{N}\sum_{n=1}^{N}\left\|\frac{1}{M}\sum_{m=1}^{M}R(S(VPI_n^m,\theta_s),\theta_r)-\hat{q}_n\right\|_1 \quad (4)$$

where $\hat{q}_n$ refers to the MOS of the n-th traning omnidirectional image.

Implementation Details

We use 80% omnidirectional images of the proposed database for training and the rest 20% for testing. We take 10 times of random train-test splitting operation and report the median performance to reduce any bias. During the training stage, we utilize the weight of ResNet-50 pre-trained on ImageNet (Deng et al. 2009) for feature extraction network initialization. The weight of other network parts is randomly initialized. The proposed model is implemented with PyTorch on an NVIDIA GeForce GTX 1080 Ti machine. We adopt Adam (Kingma and Ba 2017) optimizer with weight decay 5×10^{-4} and set mini-batch size to 4 for training 100 epochs, where the learning takes roughly a half-day. The initial learning rate is set to 10^{-4} and reduced by a decay factor 0.1 for 50 epochs. The size of input viewport images is $224 \times 224 \times 3$. During the testing stage, the final prediction quality score of a test omnidirectional image is computed by averaging all the corresponding viewport images predictions.

Experiments

We first introduce the evaluation criteria, and then measure the prediction results of the proposed methods and several state-of-the-art IQA and OIQA methods on the proposed database. Finally, we carry out the performance comparison by analysing the experimental results.

Evaluation Criteria

Three standard performance criteria, including Pearson's linear correlation coefficient (PLCC), Spearman's rank order correlation coefficient (SRCC) and root mean square error (RMSE), are used to measure the prediction monotonicity and accuracy. OIQA model with better performance has higher values of PLCC and SRCC and lower RMSE values. As suggested in (VQEG 2000), the predicted scores are first mapped to subjective ratings before calculating PLCC for maximizing the correlation between them and a four-parameter logistic function is adopted.

Performance Comparison

To demonstrate the effectiveness of the proposed models, several state-of-the-art IQA and OIQA models are used for comparison, which can be classified: traditional 2D FR-IQA metrics, *i.e.*, SSIM (Wang et al. 2004), FSIM (Zhang et al. 2011), VIF (Sheikh and Bovik 2006), VSI (Zhang, Shen, and Li 2014) and DISTS (Ding et al. 2020); 2D NR-IQA metrics, *i.e.*, NIQE (Mittal, Soundararajan, and Bovik

Type	Metrics	BD			GB			GN			ST			Overall		
		PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE
FR-IQA	SSIM	0.2753	0.2470	0.7381	0.2429	0.2159	0.8373	0.2776	0.2208	0.7420	0.0560	-0.0221	0.7284	0.1333	0.0568	0.7799
	FSIM	0.3288	0.2972	0.7251	0.2506	0.2231	0.8356	0.2081	0.1809	0.7725	0.0509	-0.0189	0.7285	0.1438	0.0690	0.7788
	VIF	0.2455	0.2191	0.7443	0.2192	0.1813	0.8422	0.2552	0.2144	0.7467	0.0757	-0.0545	0.7274	0.1786	0.1086	0.7743
	VSI	0.2830	0.2494	0.7364	0.2644	0.2360	0.8325	0.2532	0.1959	0.7472	-0.0026	-0.0004	0.7295	0.0600	0.0672	0.7855
	DISTS	0.2757	-0.2471	0.7380	0.2872	-0.2576	0.8268	0.2980	-0.2666	0.7372	0.1304	-0.0976	**0.7233**	0.1660	-0.0899	0.7760
NR-IQA	NIQE	0.0556	-0.0041	0.7666	0.0224	-0.0081	0.8630	0.0956	-0.0673	0.7691	0.0462	-0.0206	0.7287	0.0473	-0.0388	0.7861
	IL-NIQE	0.0649	-0.0256	0.7663	0.0606	-0.0510	0.8616	0.0694	-0.0701	0.7705	0.0544	-0.0535	0.7291	0.0446	-0.0447	0.7862
FR-OIQA	S-PSNR	0.2949	0.2591	0.7336	0.2137	0.1877	0.8432	0.3069	0.2452	0.7350	0.0558	-0.0479	0.7284	0.1283	0.0394	0.7804
	WS-PSNR	0.2942	0.2572	0.7338	0.2116	0.1858	0.8436	0.3094	0.2498	0.7344	0.0560	-0.0483	0.7283	0.1260	0.0381	0.7807
	CPP-PSNR	0.2942	0.2572	0.7338	0.2116	0.1860	0.8436	0.3083	0.2475	0.7347	0.0561	-0.0484	0.7283	0.1270	0.0385	0.7806
	WS-SSIM	0.2293	0.2063	0.7473	0.1710	0.1440	0.8505	**0.3260**	**0.2869**	**0.7301**	0.0454	-0.0332	0.7287	0.1168	0.0983	0.7816
NR-OIQA	MC360IQA	0.2645	0.2238	0.7242	0.0696	0.1525	0.8618	0.1657	0.0761	0.7797	0.1606	0.1572	0.7554	0.1514	0.1351	0.7838
	Ours	**0.7381**	**0.7383**	**0.5066**	**0.3369**	**0.2519**	**0.8134**	0.1490	0.1433	0.7818	**0.1699**	**0.1758**	0.7542	**0.6422**	**0.6171**	**0.6079**

Table 3: Performance comparison of state-of-the-art IQA and OIQA methods on the proposed database. In each column, the best results are highlighted in bold.

2013) and IL-NIQE (Zhang, Zhang, and Bovik 2015); traditional FR-OIQA metrics *i.e.*, S-PSNR (Yu, Lakshman, and Girod 2015), WS-PSNR (Sun, Lu, and Yu 2017), CPP-PSNR (Zakharchenko, Choi, and Park 2016), WS-SSIM (Zhou et al. 2018) and a learning-based NR-OIQA metric named MC360IQA (Sun et al. 2020). We implement all compared methods on the proposed database by using the public codes from the respective authors. To make a fair comparison, we retrain the MC360IQA (Sun et al. 2020) by using the same training/testing split scheme. Table 3 summarizes the performance on the proposed database.

According to the quantitative results from the table, we have several interesting observations. First, it's obvious that all traditional 2D FR and NR IQA models fail when they are directly applied to assess the quality of the proposed database. The performance of the former is higher due to the presence of reference content. Although DISTS (Ding et al. 2020) has comparable results by means of a CNN-based multi-scale overcomplete representations, the results indicate that current 2D IQA models do not work well on OIQA and the properties of omnidirectional images are ignored. Second, compared with traditional 2D FR-IQA methods, there is no significant improvement or even a decrease for current FR-OIQA models for the overall database. This also suggests recent advanced 2D IQA methods may be introduced into VR researches rather than just relying on existing standard metrics. Third, the latest CNN-based MC360IQA (Sun et al. 2020) model performs poorly on the proposed database. Nevertheless, as observed in Table 3, the proposed model achieves significant performance improvements from the overall database or individual distortions compared to state-of-the-art OIQA methods. This may be benefited from the established content-aware network framework and full use of human's information including viewing conditions and behaviors. The proposed model cannot efficiently handle various distortion and it is still challenging to cover each non-uniform distortions for OIQA. In brief, the experimental results on the proposed database show that the proposed method achieves the best prediction performance among the existing 2D IQA and OIQA methods, which further proves the effectiveness of the proposed method and the complexity of the proposed database.

Ablation Study

We test the performance of the proposed method with different pooling ways (Li, Jiang, and Ming 2019). From Table 4, we find that the proposed model using average pooling performs better than that using temporal pooling. This is reasonable and explicable, since the viewports we extract are obtained by sampling directly from omnidirectional images, rather than the user viewing behavior (*i.e.* scanpaths). It would be unreasonable that we use the temporal pooling strategy for quality computation. Therefore, we use the average pooling method in the proposed metric.

Strategies	SRCC	PLCC	RMSE
AP	0.6422	0.6171	0.6079
TP	0.5317	0.4904	0.6716

Table 4: Performance comparison between the proposed methods with different pooling strategies. AP: average pooling. TP: temporal pooling.

Conclusion

In this paper, we conduct a comprehensive study on perceptual quality assessment of omnidirectional images from both subjective and objective perspectives. Specifically, we construct a large-scale OIQ database, where we find that the interaction of both distortion type and starting point has a significant impact on the perception of image quality. Furthermore, we propose a new OIQA model that includes a multi-scale feature extraction module and a perceptual quality prediction module, emphasizing the incorporation of viewing conditions into the process of quality assessment. Experimental results on the proposed database validate the promising performance of the proposed method compared with state-of-the-art methods. However, experiments also show that there is still plenty of room for improvement. In the future, more important aspects such as rational viewport extraction, authentic scanpath prediction, and advanced subjective-inspired temporal pooling strategies will be considered to develop objective models with high robustness and efficiency.

Acknowledgements

This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0100601.

References

BT, R. I.-R. 2002. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*.

Chen, M.; Jin, Y.; Goodall, T.; Yu, X.; and Bovik, A. C. 2020. Study of 3D virtual reality picture quality. *IEEE Journal of Selected Topics in Signal Processing*, 14(1): 89–102.

Chen, S.; Zhang, Y.; Li, Y.; Chen, Z.; and Wang, Z. 2018. Spherical structural similarity index for objective omnidirectional video quality assessment. In *IEEE International Conference on Multimedia and Expo*, 1–6.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Li, F.-F. 2009. ImageNet: A large-scale hierarchical image database. In *IEEE Conferenc on Computer Vision and Pattern Recognition*, 248–255.

Ding, K.; Ma, K.; Wang, S.; and Simoncelli, E. P. 2020. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. To appear.

Duan, H.; Zhai, G.; Min, X.; Zhu, Y.; Fang, Y.; and Yang, X. 2018. Perceptual quality assessment of omnidirectional images. In *IEEE International Symposium on Circuits and Systems*, 1–5.

Fang, Y.; Sui, X.; Wang, J.; Yan, J.; Lei, J.; and Le Callet, P. 2021. Perceptual quality assessment for asymmetrically distorted stereoscopic video by temporal binocular rivalry. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8): 3010–3024.

Fang, Y.; Zhang, C.; Huang, H.; and Lei, J. 2019. Visual attention prediction for stereoscopic video by multi-module fully convolutional network. *IEEE Transactions on Image Processing*, 28(11): 5253–5265.

Fang, Y.; Zhang, X.; and Lmamoglu, N. 2018. A novel superpixel-based saliency detection model for 360-degree images. *Signal Processing: Image Communication*, 69: 1–7.

Hasler, D.; and Suesstrunk, S. E. 2003. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, volume 5007, 87–95.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *IEEE Conferenc on Computer Vision and Pattern Recognition*, 770–778.

ITU-T RECOMMENDATION, P. 1999. Subjective video quality assessment methods for multimedia applications. *International Telecommunication Union*.

Jiang, H.; Jiang, G.; Yu, M.; Zhang, Y.; Yang, Y.; Peng, Z.; Chen, F.; and Zhang, Q. 2021. Cubemap-Based Perception-Driven Blind Quality Assessment for 360-degree Images. *IEEE Transactions on Image Processing*, 30: 2364–2377.

Kim, H. G.; Lim, H.-T.; and Ro, Y. M. 2020. Deep virtual reality image quality assessment with human perception guider for omnidirectional image. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(4): 917–928.

Kingma, D. P.; and Ba, J. 2017. Adam: A method for stochastic optimization. arXiv:1412.6980.

Li, D.; Jiang, T.; and Ming, J. 2019. Quality assessment of in-the-wild videos. In *ACM International Conference on Multimedia*, 2351–2359.

Lim, H.-T.; Kim, H. G.; and Ra, Y. M. 2018. VR IQA NET: Deep Virtual Reality Image Quality Assessment Using Adversarial Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6737–6741.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2013. Making a completely blind image quality analyzer. *IEEE Signal Processing Letters*, 20(3): 209–212.

Sheikh, H. R.; and Bovik, A. C. 2006. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2): 430–444.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blind assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3667–3676.

Sui, X.; Ma, K.; Yao, Y.; and Fang, Y. 2021. Perceptual quality assessment of omnidirectional images as moving camera videos. *IEEE Transactions on Visualization and Computer Graphics*. To appear.

Sun, W.; Gu, K.; Zhai, G.; Ma, S.; Lin, W.; and Le Calle, P. 2017. CVIQD: Subjective quality evaluation of compressed virtual reality images. In *IEEE International Conference on Image Processing*, 3450–3454.

Sun, W.; Min, X.; Zhai, G.; Gu, K.; Duan, H.; and Ma, S. 2020. MC360IQA: A multi-channel cnn for blind 360-degree image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 14(1): 64–77.

Sun, Y.; Lu, A.; and Yu, L. 2017. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, 24(9): 1408–1412.

Upenik, E.; Řeřábek, M.; and Ebrahimi, T. 2016. Testbed for subjective evaluation of omnidirectional visual content. In *Picture Coding Symposium*, 1–5.

VQEG. 2000. Final report from the video quality experts group on the validation of objective models of video quality assessment. http://www.vqeg.org.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wang, Z.; and Rehman, A. 2017. Begin with the end in mind: A unified end-to-end quality-of-experience monitoring, optimization and management framework. In *SMPTE Annual Technical Conference and Exhibition*, 1–11.

Xu, J.; Lin, C.; Zhou, W.; and Chen, Z. 2018. Subjective quality assessment of stereoscopic omnidirectional image. In *Pacific Rim Conference on Multimedia*, 589–599.

Xu, J.; Luo, Z.; Zhou, W.; Zhang, W.; and Chen, Z. 2019. Quality assessment of stereoscopic 360-degree images from multi-viewports. In *Picture Coding Symposium*, 1–5.

Xu, J.; Zhou, W.; and Chen, Z. 2021. Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5): 1724–1737.

Yan, J.; Zhong, Y.; Fang, Y.; Wang, Z.; and Ma, K. 2021. Exposing semantic segmentation failures via maximum discrepancy competition. *International Journal of Computer Vision*, 129(5): 1768–1786.

Yang, L.; Xu, M.; Deng, X.; and Feng, B. 2021. Spatial attention-based non-reference perceptual quality prediction network for omnidirectional images. In *IEEE International Conference on Multimedia and Expo*, 1–6.

Ye, Y.; Alshina, E.; and Boyce, J. 2017. JVET-G1003: Algorithm description of projection format conversion and video quality metrics in 360lib version 4. *Joint Video Exploration Team*.

Yu, M.; Lakshman, H.; and Girod, B. 2015. A framework to evaluate omnidirectional video coding schemes. In *IEEE International Symposium on Mixed and Augmented Reality*, 31–36.

Zakharchenko, V.; Choi, K. P.; and Park, J. H. 2016. Quality metric for spherical panoramic video. In *Optics and Photonics for Information Processing X*, volume 9970, 99700C.

Zhang, L.; Shen, Y.; and Li, H. 2014. VSI: A visual saliency-index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10): 4270–4281.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A feature-enriched completely blind image quality evaluator. *IEEE Signal Processing Letters*, 24(8): 2579–2591.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. FSIM: A feature similarit index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8): 2378–2386.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2020. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 36–47.

Zhang, W.; Ma, K.; Zhai, G.; and Yang, X. 2021. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30: 3474–3486.

Zhou, W.; Xu, J.; Jiang, Q.; and Chen, Z. 2021a. No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness. *IEEE Transactions on Circuits and Systems for Video Technology*. To appear.

Zhou, Y.; Sun, Y.; Li, L.; Gu, K.; and Fang, Y. 2021b. Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network. *IEEE Transactions on Circuits and Systems for Video Technology*. To appear.

Zhou, Y.; Yu, M.; Ma, H.; Shao, H.; and Jiang, G. 2018. Weighted-to-spherically-uniform SSIM objective quality evaluation for panoramic video. In *2018 14th IEEE International Conference on Signal Processing*, 54–57.