Augmented Reality Image Quality Assessment Based on Visual Confusion Theory

Huiyu Duan[1*], Lantu Guo[2,3*], Wei Sun[1], Xiongkuo Min[1], Li Chen[1], and Guangtao Zhai[1§]
[1]*Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China*
[2]*Beijing Institute of Technology, Beijing, China*
[3]*China Research Institute of Radiowave Propagation, Qingdao, China*
{huiyuduan, sunguwei, minxiongkuo, hilichen, zhaiguangtao}@sjtu.edu.cn, guolantu@163.com

2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) | 978-1-6654-6901-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/BMSB55706.2022.9828671

***Abstract*—With the development of multimedia technology, Augmented Reality (AR) has become a promising next-generation mobile platform. In order to acquire better AR mobile streaming experience, it is significant to study the evaluation of Quality of Experience (QoE) in AR. The primary value of AR is to promote the fusion of digital contents and real-world environments, however, studies on how this fusion will influence the Quality of Experience (QoE) of these two components are lacking. To achieve better QoE of AR, whose two layers are influenced by each other, it is important to evaluate its perceptual quality first. In this paper, we consider AR technologies as the *superimposition* of virtual scenes and real scenes, and introduce *visual confusion* as its basic theory. We first establish an ARIQA database to better simulate the real AR application scenarios, which contains 20 AR reference images, 20 background (BG) reference images, and 560 distorted images generated from AR and BG references by mixing reference images in pairs, as well as the correspondingly collected subjective quality ratings. We also design three types of full-reference (FR) IQA metrics to study whether we should consider the visual confusion when designing corresponding IQA algorithms. An ARIQA metric is finally proposed for better evaluating the perceptual quality of AR images. The dataset, benchmark study and proposed metric will be released to facilitate the future studies related to AR QoE assessment and AR communication.**

***Index Terms*—Subjective evaluation techniques, objective evaluation techniques, quality of Experience, Augmented Reality (AR), visual confusion**

I. Introduction

With the evolution of multimedia technology, the next-generation display technologies aim at revolutionizing the way of interactions between users and their surrounding environment rather than limiting to flat panels that are just placed in front of users (*i.e.*, mobile phone, computer, *etc.*) [1], [2]. These technologies, including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), *etc.*, have been developing rapidly in recent years. Among them, AR pursues high-quality see-through performance and enriches the real world by superimposing digital contents on it, which is promising to become the next-generation mobile platform. With advanced experience, AR shows great potential in several attractive application scenarios, including but not limited to communication, entertainment, health care, education, engineering design, *etc.*

* Equal contribution.
§ Corresponding author.

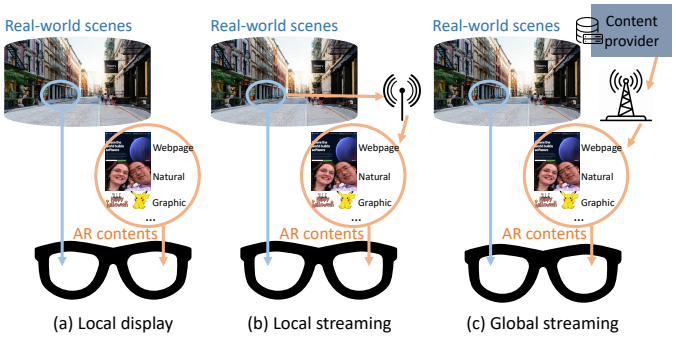

Fig. 1. Augmented Reality image quality assessment can be used for optimizing local display, local streaming (AR contents provided by local real-world), and global streaming (AR contents provided by remote server). Note that real-world scenes and AR contents can be displayed binocularly (in this figure) or monocularly (in Fig. 2), we mainly consider monocular visual confusion in this paper.

On account of the complex application scenes, it is important to study the perceptual Quality of Experience (QoE) of AR, which can help better optimize local display, local streaming, and global streaming of mobile AR as shown in Fig. 1. Lately, some works have been presented to study the quality effects of typical degradations that affect digital contents in AR [3], [4]. These studies have performed subjective/objective tests on screen displays via showing videos of 3D meshes or point clouds with various distortions. Moreover, with the development of Head Mounted Displays (HMDs) for AR applications, some studies have considered evaluating the QoE of 3D objects using these devices. However, all these studies only focus on the degradations of geometry and texture of 3D meshes and point clouds inside AR, *e.g.,* noise, compression *etc.*, their see-through scenes are either blank or simple texture, or even without see-through scenes (opaque images/objects). The studies discussing the relationship between augmented view and see-through view are lacking.

To address the above issues, in this paper, we consider AR technology as the *superimposition* of digital contents and see-through contents, and introduce ***visual confusion*** [5], [6] as its basic theory. Although the digital contents can enrich the information of real world scene and bring much convenience, the superimposition may degrade the quality of AR contents or occlude the real world scene. Thus, it is important to study the IQA for AR contents and further use it for improving the

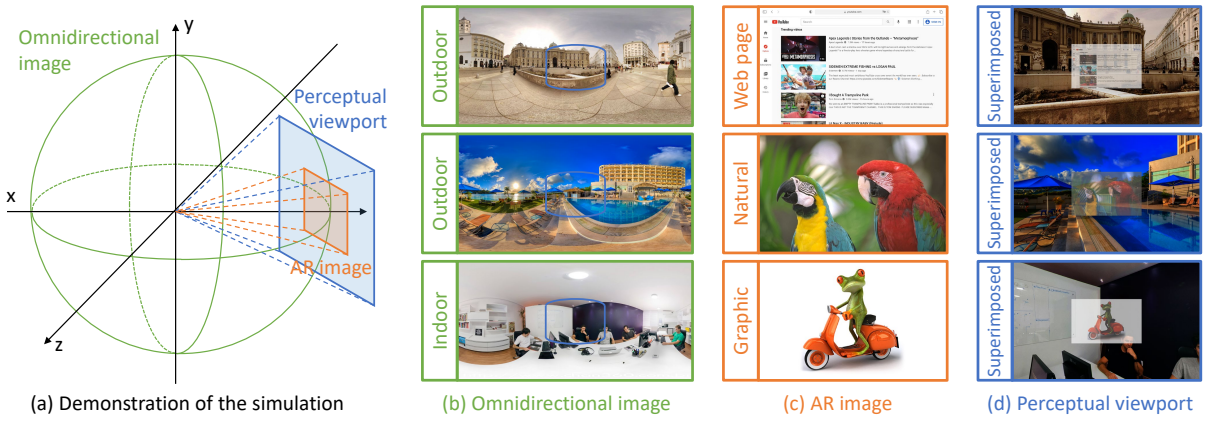

Fig. 2. The illustration of the AR simulation in VR environment. (a) The demonstration of the relationship between the omnidirectional image, the AR image, and the perceptual viewport image. (b) The omnidirectional images are used as the background scenes, which include outdoor and indoor scenarios. (c) The AR images are composed of three types of content including web page images, natural images, and graphic images. (d) The perceptual viewport images are generated by superimposing the AR images on the omnidirectional images (here $\lambda = 0.58$). Note that the perceptual viewports of the subjects are changed dynamically with the head movement, however, the relative positional relationship between the omnidirectional image and the AR image is fixed.

QoE of AR. To this end, in this work, we conduct the first subjective and objective ARIQA study based on the visual confusion theory.

An Augmented Reality Image Quality Assessment (ARIQA) dataset is established to make up for the absence of relevant research. An intuitive way to conduct subjective AR experiment is wearing AR devices in various environments and then collecting subjective scores. However, this way suffers from uncontrollable experimental environments and limited experimental scenarios. Moreover, A big TV screen cannot provide immersive experience and enough field-of-view for the background image. Thus, we innovatively propose to conduct the ARIQA study in VR environment. Specifically, we first collect 20 raw omnidirectional images as background images, and 20 common images (including graphic images, natural images, and webpage images) as reference images. Then 20 background images (omnidirectional images) and 20 reference AR images are randomly combined in 20 background-AR (B-A) pairs. Four mixing levels are introduced as the visual confusion distortion for these 20 pairs during the experiment. Besides the visual confusion distortion as mentioned above, we further introduce three types of distortions to AR images including JPEG compression, image scaling and contrast adjustment to AR contents, and each of these distortions has two levels. Finally, we generate 560 B-A pairs as the test stimuli ($20\ scenarios \times 7\ levels \times 4\ mixing\ values$).

We also design three types of objective AR-IQA benchmark models, which can be differentiated according to the inputs of the classical IQA models, to study whether and how the visual confusion should be considered when designing corresponding IQA metrics. Based on the ARIQA dataset and the benchmark models, we further analyze several visual characteristics of visual confusion and propose an attention based deep feature fusion method towards better evaluating the quality of superimposed images. Specifically, the attention based deep feature fusion model is established based on LPIPS [7], we subtract the DNN features extracted from a superimposed image and the corresponding AR image, as well as the DNN features extracted from the superimposed image and the background image respectively, as two feature distance vectors. Then these two feature distance vectors are fed into a channel attention module and a spatial attention module to further refine the feature vectors. For training efficiency, the spatial attention module is a modified and frozen saliency prediction model. Finally, these two feature distance vectors are fed into an average pooling layer and subtracted to get the final quality score. A specialized learning strategy is also proposed. Specifically, two superimposed images and corresponding AR, background images are fed into two aforementioned attention based deep feature fusion model respectively to obtain two quality scores of these two contents, and by calculating the loss function between the subtraction of these two scores and the ground-truth subtraction, the relative quality comparison can be learned by the network.

The rest of this paper is organized as follows. Section II introduces the subjective ARIQA methodology. In Section III, we presents the benchmark methods and our proposed ARIQA model. The results are given in Section IV. Section V concludes the whole paper.

II. Subjective ARIQA Methodology

A. Subjective Experiment Methodology.

An intuitive way to conduct subjective AR experiment is wearing AR devices in various environments and then collecting subjective scores. However, this way suffers from uncontrollable experimental environments and limited experimental scenarios [8], *e.g.*, the head movement may cause different collected background images for different users, and it is hard to introduce various background scenarios in lab environment. Therefore, we adopt the method of conducting subjective AR-IQA studies in VR environment for controllable experimental environments and diverse experimental scenarios.

Fig. 2 illustrates the methodology of the subjective experiment in this ARIQA study. First of all, 20 omnidirectional

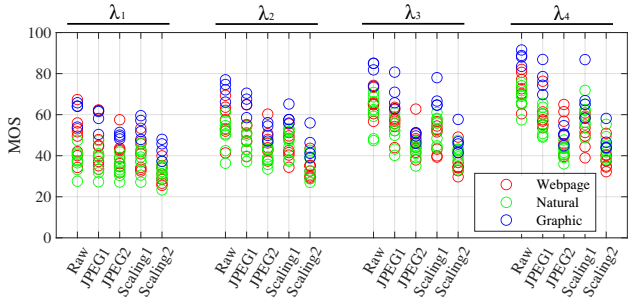

Fig. 3. Distribution of MOS values of raw images, JPEG compressed images, rescaled images superimposed on the omnidirectional backgrounds with different mixing values. The mixing values λ_1, λ_2, λ_3, λ_4 are equal to 0.26, 0.42, 0.58, 0.74, respectively.

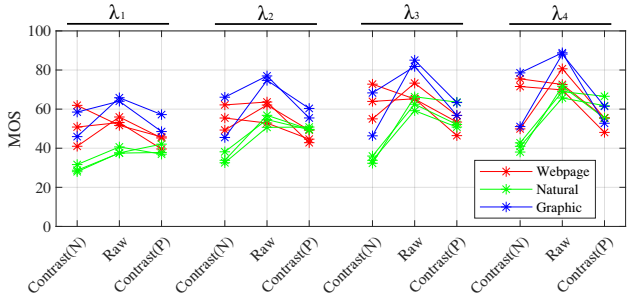

Fig. 4. Samples of MOS values of raw images, contrast adjusted images superimposed on the omnidirectional backgrounds with different mixing values. "N" denotes negative gamma transfer, "P" represents positive gamma transfer.

images are collected as the background scenes including 10 indoor scenarios and 10 outdoor scenarios. Considering the real applications of AR, we further collect 20 images as the reference AR contents, which include 8 web page images, 8 natural images, and 4 graphic images. The resolution of all raw AR images is 1440×900. We generate a much larger set of distorted AR contents by applying quality degradation processes that would occur in AR applications. Three distortion types including image compression, image scaling, image contrast adjustment, are introduced as follows. 1) *JPEG compression*. We set the quality level of the JPEG compression at the two levels with quality parameters 7 and 3. 2) *Image scaling*. We create distorted images by downsclaing original images to 1/5 and 1/10 of the original resolution, then spatially upscaling them back to the original resolution. 3) *Image contrast adjustment.* We use the gamma transfer function [9] to adjust the contrast, which is defined as $y = [x \cdot 255^{((1/n)-1)}]^n$, where $n = [1/4, 4]$ ($n < 1$ is negative gamma transfer, $n > 1$ is positive gamma transfer). Hence, for each AR image, we generate 6 degraded images.

For simulating AR scenarios, we first randomly match the 20 AR images and 20 omnidirectional images in pairs to generate 20 scenarios. Hence, for each omnidirectional image, we have 7 AR images superimposed on it (1 reference image + 6 distorted images). During the experiment, the perceptual viewport can be formulated as:

$$I_S = \lambda \circ I_A + (1 - \lambda) \circ I_O \ , \tag{1}$$

where I_S denotes the perceptual viewport, *i.e.*, the superimposed image, I_A represents the AR image, I_O indicates the omnidirectional image, and $\lambda \in [0.26, 0.42, 0.58, 0.74]$ denotes the mixing value used in the experiment, *i.e.*, we have four superimposing levels in this subjective experiment. Overall, 560 experimental stimuli are generated for conducting the subjective experiment (20 $scenarios \times 7$ $levels \times 4$ $mixing$ $values$). As demonstrated in Fig. 2 (a), the omnidirectional image is displayed in 360 degrees as the background scenarios, the AR image is superimposed on the omnidirectional image which is perceived as the perceptual viewport. Fig. 2 (b), (c) and (d) present the examples of the omnidirectional images, the AR images, and the perceptual viewport images, respectively.

A total of 23 subjects participate in the experiment. Since the experiment is conducted under VR-HMD environment, the single-stimulus (SS) strategy is adopted to collect the subjective quality ratings of AR images. A 10-point numerical categorical rating method is used to facilitate the subjective rating in HMD [10]. We use HTC VIVE Pro Eye [11] as the HMD on account of its excellent graphics display technology and high precision tracking ability. During the formal test, all 560 experimental stimuli are displayed in a random order for each subject. We then process the collected subjective scores to obtain the mean opinion scores (MOSs).

B. Subjective Data Analysis.

We analyze the distribution of MOS values across different mixing values and various distortions. Fig. 3 shows the MOS distribution of the images with the degradations of JPEG compression and image scaling under different mixing values. We notice that as the λ value increases, the MOS value also shows an overall upward trend, of which the reason is apparent since larger λ value means clearer AR content. Moreover, for the superimposed AR images with JPEG compression and scaling, we notice that when the mixing value λ is relatively smaller, the MOSs of these images are closer to that of superimposed raw images, though the overall MOSs are smaller than that of the larger λ values. Fig. 4 plots several examples of the MOS values of raw images and contrast adjusted images superimposed on the omnidirectional backgrounds with different mixing values, which shows that appropriate contrast adjustment may even improve the perceptual quality of AR contents.

III. Objective ARIQA Model

A. Benchmark Method

Three variants are introduced in the benchmark study. We assume the background image, the AR image, as well as the mixing value are known, which can be acquired in real applications, and the superimposed image can be correspondingly calculated. Let I_{A_D} denotes the AR image with distortions, I_{A_R} denotes the raw reference AR image, I_B indicates the background image, λ represents the mixing value, hence, the displayed AR image I_A and the perceptual viewport image (superimposed image) I_S can be correspondingly expressed as:

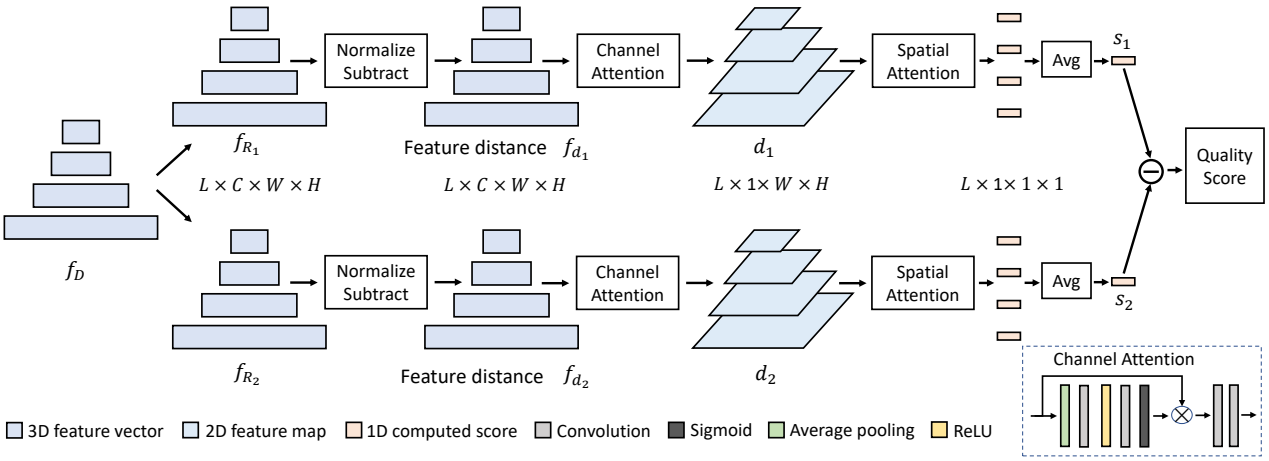

Fig. 5. Attention based deep feature fusion. x_D, x_{R_1}, and x_{R_2} are three extracted feature vectors, respectively. We first compute the feature distance between the corresponding feature layers of the distorted image and two reference images. Then each feature distance vector is fed into a specially designed channel attention module and a one dimensional feature map is output. After weighting by a spatial attention operation, the predicted score is computed.

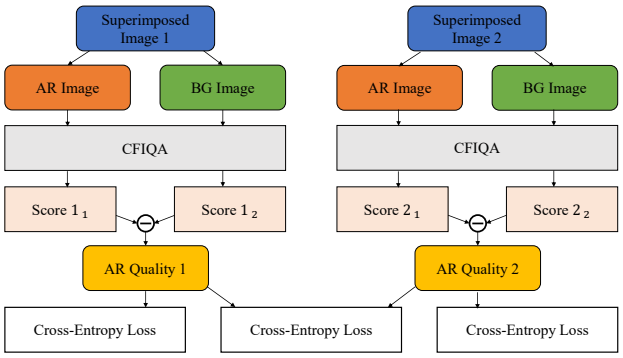

Fig. 6. The framework of the proposed ARIQA model.

$I_A = \lambda \cdot I_{A_D}$, and $I_S = \lambda \cdot I_{A_D} + (1-\lambda) \cdot I_B$, respectively. Then, three FR-IQA variants of AR IQA metrics are defined as: Type I, the similarity between the displayed AR image I_A and the reference AR image I_{A_R}; Type II, the similarity between the perceptual viewport image I_S and the reference AR image I_{A_R}; Type III, the SVR fusion [12] of the similarity between the perceptual viewport image I_S and the reference AR image I_{A_R}, and the similarity between the perceptual viewport image I_S and the background image I_B. These three variant types can be expressed as:

$$Q_{\text{Type I}} = \text{FR}(I_A, I_{A_R}), \tag{2}$$

$$Q_{\text{Type II}} = \text{FR}(I_S, I_{A_R}), \tag{3}$$

$$Q_{\text{Type III}} = \text{SVR}(\text{FR}(I_S, I_{A_R}), \text{FR}(I_S, I_B)), \tag{4}$$

where $Q_{\text{Type I}}$, $Q_{\text{Type II}}$, and $Q_{\text{Type III}}$ denote the quality predictions of the three variants, SVR indicates the support vector regression deployment.

B. Attention Based Deep Feature Fusion Method (CFIQA)

We first introduce an attention based deep feature fusion method for confusing image quality assessment (CFIQA) as shown in Fig. 5.

Deep feature extraction and feature distance calculation. We first employ several state-of-the-art pre-trained DNNs to extract both low-level and high-level features, which include SqueezeNet [13], AlexNet [14], VGG Net [15], and ResNet [16]. Then, for a distorted image I_D and corresponding two reference images I_{R_1}, I_{R_2}, we extract feature stacks f_D, f_{R_1}, and f_{R_2} from L layers of a network $\mathcal{F}$, respectively. Then we follow the method in [7] and calculate the feature distance between the distorted image and the reference image by subtracting normalized feature stacks. This can be expressed as:

$$f_{d_i}^l = \left\| \, f_D^l - f_{R_i}^l \, \right\|_2^2 , \tag{5}$$

where $l \in [1, L]$ represents the l-th layer, $i \in \{1, 2\}$ denotes the reference category, $f_{d_i}^l$ is the calculated feature distance.

Channel attention for learning feature significance. Since the significance of each channel of the feature distance vector is uncertain for this task, it is important to learn the weights of the channels for each feature distance vector and re-organize them. We adopt a widely used channel attention [30] method as shown in Fig. 5 to learn and re-organize features. After the channel attention, two down-sampling convolutional layers are followed. The kernel size of all convolutional layers is 1. Through this manipulation, a 2D feature stack d_i which represents the distance map can be obtained for f_{d_i}, where $i \in \{1, 2\}$.

Spatial attention. Since high-level visual features such as saliency may influence the perceptual quality of visual confusion. Since it is hard to optimize spatial attention with a relatively small dataset, in this work, we calculate spatial attention by a saliency prediction method. A state-of-the-art saliency prediction method [31] is used to to calculate the spatial attention map W_i for a reference image I_{R_i}. By weighting the distance map d_i with a scaled spatial attention

TABLE I

PERFORMANCE OF THE THREE VARIANTS OF THE STATE-OF-THE-ART FR-IQA MODELS ON THE ARIQA DATASET. THE TOP 3 RESULTS OF ALL THREE VARIANTS ARE IN **BOLD** FOR EACH GROUP. THE PERFORMANCE CHANGES COMPARED TO TYPE I IN TERMS OF SRCC ARE INDICATED IN GRAY FONTS

Method	Type I				Type II				Type III			
Model \ Criteria	SRCC↑	KRCC↑	PLCC↑	RMSE↓	SRCC↑	KRCC↑	PLCC↑	RMSE↓	SRCC↑	KRCC↑	PLCC↑	RMSE↓
PSNR	0.2197	0.1485	0.2742	12.733	0.0064 (−0.2133)	0.0027	0.0592	13.217	0.3809 (+0.1612)	0.2662	0.4154	11.901
NQM [17]	0.4101	0.2813	0.4268	11.974	0.5348 (+0.1247)	0.3772	0.5550	11.014	0.5588 (+0.1487)	0.4031	0.5867	10.677
MS-SSIM [18]	0.6118	0.4414	0.6483	10.080	0.6557 (+0.0439)	0.4778	0.6609	9.9366	0.6660 (+0.0541)	0.4914	0.6741	9.6721
SSIM [19]	0.5327	0.3799	0.5551	11.013	0.5399 (+0.0072)	0.3797	0.5411	11.134	0.6090 (+0.0763)	0.4430	0.6233	10.276
IFC [20]	0.3539	0.2456	0.3294	12.501	0.5121 (+0.1582)	0.3523	0.5105	11.385	0.5090 (+0.1551)	0.3601	0.5217	11.172
VIF [21]	0.5981	0.4273	0.6366	10.211	0.6927 (+0.0946)	0.5009	0.6869	9.6218	**0.7227** (+0.1245)	**0.5351**	**0.7222**	**9.2024**
IW-MSE [22]	0.2287	0.1555	0.2966	12.644	0.2406 (+0.0119)	0.1689	0.2956	12.648	0.4126 (+0.1839)	0.2906	0.4586	11.693
IW-PSNR [22]	0.2287	0.1555	0.2998	12.631	0.2406 (+0.0119)	0.1689	0.2895	12.673	0.3559 (+0.1272)	0.2574	0.4151	11.879
IW-SSIM [22]	0.6431	0.4663	0.6532	10.026	**0.7103** (+0.0672)	**0.5267**	**0.7100**	**9.3231**	**0.7116** (+0.0685)	**0.5337**	**0.7201**	**9.0193**
FSIM [23]	0.6323	0.4546	0.6723	9.8010	0.6538 (+0.0215)	0.4716	0.6528	10.029	0.6663 (+0.0340)	0.4865	0.6774	9.5764
GSI [24]	0.4393	0.3046	0.5034	11.440	0.3788 (−0.0605)	0.2606	0.3890	12.197	0.4245 (−0.0147)	0.3056	0.4584	11.680
GMSD [25]	0.6485	0.4718	0.6759	9.7575	0.5947 (−0.0537)	0.4346	0.5959	10.633	0.6730 (+0.0245)	0.4973	0.6801	9.5815
GMSM [25]	0.6386	0.4628	0.6907	9.5745	0.5863 (−0.0523)	0.4142	0.5923	10.667	0.6294 (−0.0092)	0.4587	0.6422	10.064
PAMSE [26]	0.2162	0.1458	0.2736	12.735	0.0090 (−0.2072)	0.0048	0.0659	13.211	0.3657 (+0.1495)	0.2558	0.4093	11.941
LTG [27]	0.6592	0.4830	0.6826	9.6759	0.6469 (−0.0123)	0.4742	0.6422	10.150	0.6764 (+0.0172)	0.4998	0.6818	9.4727
VSI [28]	0.5190	0.3691	0.5926	10.665	0.6096 (+0.0906)	0.4318	0.6167	10.422	0.6321 (+0.1131)	0.4590	0.6484	10.039
LPIPS (Squeeze) [7]	0.5924	0.4326	0.6160	10.430	0.6260 (+0.0336)	0.4450	0.6251	10.334	0.6417 (+0.0494)	0.4693	0.6660	9.8086
LPIPS (Alex) [7]	0.5870	0.4273	0.6314	10.267	0.6306 (+0.0436)	0.4457	0.6352	10.226	**0.6626** (+0.0757)	0.4820	**0.6767**	**9.6071**
LPIPS (VGG) [7]	0.5436	0.3828	0.5593	10.975	0.6202 (+0.0766)	0.4426	0.6141	10.450	0.6373 (+0.0936)	0.4606	0.6475	9.9848
DISTS [29]	0.5011	0.3583	0.5280	11.244	0.5112 (+0.0101)	0.3627	0.5528	11.033	0.6334 (+0.1323)	0.4608	0.6580	9.7866
Baseline (SqueezeNet)	0.5733	0.4166	0.6096	10.496	0.6339 (+0.0606)	0.4570	0.6358	10.220	0.6272 (+0.0539)	0.4573	0.6493	9.8747
Baseline (AlexNet)	0.5273	0.3776	0.5814	10.772	0.6450 (+0.1177)	0.4690	0.6578	9.9728	0.6460 (+0.1187)	0.4768	0.6707	9.7499
Baseline (VGG-16)	0.5541	0.3908	0.5706	10.873	0.6368 (+0.0827)	0.4585	0.6372	10.204	0.6587 (+0.1046)	0.4805	0.6622	9.8906
Baseline (VGG-19)	0.5612	0.3981	0.5790	10.795	0.6561 (+0.0949)	0.4750	0.6530	10.028	0.6613 (+0.1001)	**0.4838**	0.6674	**9.6720**
Baseline (ResNet-18)	0.5438	0.3892	0.5750	10.832	0.6467 (+0.1029)	0.4678	0.6451	10.117	0.6485 (+0.1047)	0.4779	0.6702	9.7504
Baseline (ResNet-34)	0.5426	0.3862	0.5771	10.813	0.6603 (+0.1177)	0.4782	0.6660	9.8765	**0.6710** (+0.1284)	**0.4959**	**0.6903**	**9.4826**
Baseline (ResNet-50)	0.5753	0.4113	0.5977	10.615	0.6510 (+0.0757)	0.4688	0.6464	10.102	**0.6619** (+0.0866)	**0.4831**	**0.6736**	9.7163

TABLE II

PERFORMANCE OF FOUR TRAINABLE MODELS.

Model \ Criteria	SRCC↑	KRCC↑	PLCC↑	RMSE↓
modified LPIPS [7]	0.7624	0.5756	0.7591	8.6935
CFIQA (Ours)	0.7787	0.5863	0.7695	8.5484
ARIQA (Ours)	0.7902	0.5967	0.7824	8.3295
ARIQA+ (Ours)	**0.8124**	**0.6184**	**0.8136**	**7.8018**

map W_i, the final quality score can be predicted as:

$$s_i = \mathtt{Avg}_l\left(\frac{\sum_{h,w} W^l_{i\,hw} \odot d^l_{i\,hw}}{\sum_{h,w} W^l_{i\,hw}}\right) . \tag{6}$$

C. ARIQA Model

Based on the aforementioned CFIQA, we further improve the learning strategy of CFIQA by comparing the quality of two homologous superimposed images as demonstrated in Figure 6. Considering the effectiveness of the training objectives of the LPIPS [7], during the training process, two pathways are introduced to ARIQA for comparing the perceptual quality of different distorted images of the one AR and background reference pair. The edges of objects can help identify their categories [32]. However, when two images are superimposed together, the intersection of the edges of two image layers may strongly influence the perceptual quality. Therefore, we further extract the features from an edge detection model [32] and concatenate them with the features extracted from one of the aforementioned classification backbones as an enhanced model, which is named ARIQA+.

IV. EXPERIMENTAL VALIDATION

Benchmark experiments. In terms of our ARIQA dataset, the background image I_B (*i.e.*, viewport of the omnidirectional image I_O) and the superimposed image I_S are captured in Unity, then the benchmark results are calculated using the aforementioned methods. Table I presents the performance of the three benchmark AR-IQA metric variants derived from the state-of-the-art FR-IQA models on the ARIQA dataset. Comparing Type I and Type II, we notice that for most FR-IQA metrics, using superimposed images as distorted images can improve the performance of the algorithm. In addition, as shown in the comparison between Type III and Type I, when superimposed images, AR images, as well as background images are jointly considered, the performance of almost all FR-IQA metrics can be further improved.

ARIQA model performance. We conduct a five-fold cross validation experiment on the ARIQA dataset. For each fold, we split the 560 samples into 280 training samples and 280 testing samples without scene repeating, *i.e.*, 280 training samples and 280 testing samples corresponding to different 10 AR/BG pairs, respectively. For fair comparison, we further re-train the LPIPS and CFIQA models only using AR image as the reference image, which is similar to the concept of Type II described above. Table II shows the averaged performance of these four models after five-fold cross validation. It can be observed that the ARIQA model achieves better performance than the LPIPS model and the CFIQA model, and the ARIQA+ achieves the best performance compared to other models.

V. Conclusion

In this paper, we discuss visual confusion theory underlying AR technologies, and conduct a subjective ARIQA study and an objective ARIQA study based on the visual confusion theory. To better study the IQA problem of AR, we first build an augmented reality image quality assessment (ARIQA) dataset, and conduct a subjective image quality assessment study based on it. Three benchmark models are presented for this problem. An ARIQA model is also proposed for better evaluating the perceptual quality of AR images. The results show that it is beneficial to consider visual confusion when designing IQA models for AR, and our proposed ARIQA model achieves better performance compared to other state-of-the-art methods. We hope this work can help other researchers to have a better understanding of the visual confusion mechanism underlying AR technology and can contribute to design and optimize AR broadband/broadcast systems.

Acknowledgment

This work was supported by the NSFC 61831015, 61901260, National Key R&D Program of China 2021YFE0206700.

References

[1] O. Cakmakci and J. Rolland, "Head-worn displays: a review," *Journal of display technology*, vol. 2, no. 3, pp. 199–216, 2006.

[2] T. Zhan, K. Yin, J. Xiong, Z. He, and S.-T. Wu, "Augmented reality and virtual reality displays: Perspectives and challenges," *Iscience*, p. 101397, 2020.

[3] J. Zhang, W. Huang, X. Zhu, and J.-N. Hwang, "A subjective quality evaluation for 3d point cloud models," in *Proceedings of the International Conference on Audio, Language and Image Processing*, 2014, pp. 827–831.

[4] E. Zerman, P. Gao, C. Ozcinar, and A. Smolic, "Subjective and objective quality assessment for volumetric video compression," *Electronic Imaging*, vol. 2019, no. 10, pp. 323–1, 2019.

[5] R. L. Woods, R. G. Giorgi, E. L. Berson, and E. Peli, "Extended wearing trial of trifield lens device for 'tunnel vision'," *Ophthalmic and physiological optics*, vol. 30, no. 3, pp. 240–252, 2010.

[6] E. Peli and J.-H. Jung, "Multiplexing prisms for field expansion," *Optometry and vision science: official publication of the American Academy of Optometry*, vol. 94, no. 8, p. 817, 2017.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[8] J. Gutiérrez, T. Vigier, and P. L. Callet, "Quality evaluation of 3d objects in mixed reality for different lighting conditions," *Electronic Imaging*, vol. 2020, no. 11, pp. 128–1, 2020.

[9] K. Gu, G. Zhai, W. Lin, and M. Liu, "The analysis of image contrast: From quality assessment to automatic enhancement," *IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 284–297, 2015.

[10] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in *2018 IEEE international symposium on circuits and systems (ISCAS)*, 2018, pp. 1–5.

[11] *HTC VIVE Pro Eye*, https://www.vive.com/us/product/vive-pro-eye/overview/.

[12] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[13] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[17] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Transactions on Image Processing (TIP)*, vol. 9, no. 4, pp. 636–650, 2000.

[18] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proceedings of the Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, 2003, pp. 1398–1402.

[19] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.

[20] H. R. Sheikh, A. C. Bovik, and G. De Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing (TIP)*, vol. 14, no. 12, pp. 2117–2128, 2005.

[21] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing (TIP)*, vol. 15, no. 2, pp. 430–444, 2006.

[22] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 20, no. 5, pp. 1185–1198, 2010.

[23] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 20, no. 8, pp. 2378–2386, 2011.

[24] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Transactions on Image Processing (TIP)*, vol. 21, no. 4, pp. 1500–1512, 2011.

[25] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing (TIP)*, vol. 23, no. 2, pp. 684–695, 2013.

[26] W. Xue, X. Mou, L. Zhang, and X. Feng, "Perceptual fidelity aware mean squared error," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 705–712.

[27] K. Gu, G. Zhai, X. Yang, and W. Zhang, "An efficient color image quality metric with local-tuned-global model," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 506–510.

[28] L. Zhang, Y. Shen, and H. Li, "Vsi: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 23, no. 10, pp. 4270–4281, 2014.

[29] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *arXiv preprint arXiv:2004.07728*, 2020.

[30] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7132–7141.

[31] R. Droste, J. Jiao, and J. A. Noble, "Unified image and video saliency modeling," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 419–435.

[32] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 41, no. 8, pp. 1939–1946, 2019.