Confusing Image Quality Assessment: Toward Better Augmented Reality Experience

Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, Patrick Le Callet

Confusing Image Quality Assessment: Towards Better Augmented Reality Experience

Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, *Senior Member, IEEE*, Xiaokang Yang, *Fellow, IEEE*, and Patrick Le Callet, *Fellow, IEEE*

***Abstract*—With the development of multimedia technology, Augmented Reality (AR) has become a promising next-generation mobile platform. 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 technology as the *superimposition* of virtual scenes and real scenes, and introduce *visual confusion* as its basic theory. A more general problem is first proposed, which is evaluating the perceptual quality of superimposed images, *i.e.,* confusing image quality assessment. A ConFusing Image Quality Assessment (CFIQA) database is established, which includes 600 reference images and 300 distorted images generated by mixing reference images in pairs. Then a subjective quality perception experiment is conducted towards attaining a better understanding of how humans perceive the confusing images. Based on the CFIQA database, several benchmark models and a specifically designed CFIQA model are proposed for solving this problem. Experimental results show that the proposed CFIQA model achieves state-of-the-art performance compared to other benchmark models. Moreover, an extended ARIQA study is further conducted based on the CFIQA study. We 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, as well as the correspondingly collected subjective quality ratings. Three types of full-reference (FR) IQA benchmark variants are designed 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. Experimental results demonstrate the good generalization ability of the CFIQA model and the state-of-the-art performance of the ARIQA model. The databases, benchmark models, and proposed metrics are available at: https://github.com/DuanHuiyu/ARIQA.**

***Index Terms*—Augmented Reality (AR), visual confusion, image quality assessment, quality of experience (QoE).**

Manuscript received October 6, 2021; revised June 2, 2022 and September 12, 2022; accepted October 19, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62225112, Grant 61831015, Grant 62271312, Grant 61901260, and Grant 62101326, in part by the National Key R&D Program of China 2021YFE0206700, in part by the Shanghai Pujiang Program under Grant 22PJ1407400, and in part by the China Postdoctoral Science Foundation 2022M712090. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chaker Larabi. *(Corresponding authors: Xiongkuo Min; Guangtao Zhai.)*

Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, and Xiaokang Yang are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: huiyuduan@sjtu.edu.cn; minxiongkuo@sjtu.edu.cn; zyc420@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn; xkyang@sjtu.edu.cn).

Patrick Le Callet is with the Polytech Nantes, Université de Nantes, 44306 Nantes, France (e-mail: patrick.lecallet@univ-nantes.fr).

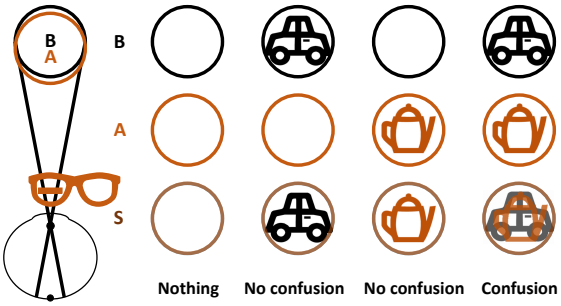


Fig. 1. Visual confusion theory of AR. "B" denotes the background view, *i.e.,* see-through view. "A" represents the augmented view. "S" implies the superimposition of the background view (B) and augmented view (A). "A" and "B" will influence the perceptual quality of each other.

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 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.*

On account of the complex application scenes, it is important to consider the perceptual Quality of Experience (QoE) of AR, which includes measuring the perceptual quality and better improving the experience of AR. Lately, some works have been presented to study the quality effects of typical degradations that affect digital contents in AR [3]–[7]. These studies have performed subjective/objective tests on screen displays 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. For instance, Alexiou *et al.* [8] have studied geometry degradations of point clouds and have conducted a subjective quality assessment study in a MR HMD system. Zhang *et al.* [9] have conducted a study towards the QoE model of AR applications using Microsoft HoloLens, which mainly focused on the perceptual factors related to the usability of AR systems. Gutierrez *et al.* [10] have proposed several

guidelines and recommendations for subjective testing of QoE in AR scenarios. 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 views and see-through views are lacking.

To address the above issues, in this paper, we consider AR technology as the *superimposition* of digital contents and see-through scenes, and introduce ***visual confusion*** [11], [12] as its basic theory. Fig. 1 demonstrates the concept of the visual confusion in AR. "B", "A" and "S" in Fig. 1 represent the background (BG) view, augmented view and superimposed view, respectively. If both "B" and "A" are views with blank/simple textures, there is nothing important in the superimposed view. If one of "B" or "A" has a complex texture, but another view has a simple texture, there is also no confusion in the superimposed view. If both "B" and "A" have complex textures, visual confusion is introduced. We assume that without introducing specific distortions that has been widely studied in the current QoE studies, visual confusion itself is a type of distortion, and it significantly influences the AR QoE. Thus, we argue that it is important to study the assessment of the visual confusion towards better improving the QoE of AR. Note that it does not mean that no confusion is better than having confusion, since the objective of AR is to promote the fusion between the virtual world and the real world. Instead, the balance between them is more important.

To this end, in this work, we first propose a more general problem, which is evaluating the perceptual quality of visual confusion. A ConFusing Image Quality Assessment (CFIQA) database is established to make up for the absence of relevant research. Specifically, we first collect 600 reference images and mix them in pairs, which generates 300 distorted images. We then design and conduct a comprehensive subjective quality assessment experiment among 17 subjects, which produces 600 mean opinion scores (MOSs), *i.e.*, for each distorted image, two MOSs are obtained for two references respectively. The distorted and reference images, as well as subjective quality ratings together constitute the ConFusing Image Quality Assessment (CFIQA) database. Based on this database, several visual characteristics of visual confusion are analyzed and summarized. Then several benchmark models and a specifically designed attention based deep feature fusion model termed CFIQA are proposed for solving this problem. Experimental results show that our proposed CFIQA model achieves better performance compared to other benchmark models.

Moreover, considering the field-of-view (FOV) of the AR image and the background image are usually different in real application scenarios, we further establish an ARIQA database for better understanding the perception of visual confusion in real world. The ARIQA database is comprised of 20 raw AR images, 20 background images, and 560 distorted versions produced from them, each of which is quality-rated by 23 subjects. Besides the visual confusion distortion as mentioned above, we further introduce three types of distortions: JPEG compression, image scaling and contrast adjustment to AR contents. Four levels of the visual confusion distortion are applied to mix the AR images and the background images. Two levels of other types of distortions are applied to the AR contents. To better simulate the real AR scenarios and control the experimental environment, the ARIQA experiment is conducted in VR environment. We also design three types of objective 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. An ARIQA model designed based on the CFIQA model is then proposed to better evaluate the perceptual quality of AR images. Experimental results demonstrate that our CFIQA model also achieves good generalization ability on the ARIQA database, and the proposed ARIQA model achieves the best performance compared to other methods.

Overall, the main contributions of this paper are summarized as follows. (i) We discuss the visual confusion theory of AR and argue that evaluating the visual confusion is one of the most important problem of evaluating the QoE of AR. (ii) We establish the first ConFusing IQA (CFIQA) database, which can facilitate further objective visual confusion assessment studies. (iii) To better simulate the real application scenarios, we establish an ARIQA database and conduct a subjective quality assessment experiment in VR environment. (iv) A CFIQA model and an ARIQA model are proposed for better evaluating the perceptual quality in these two application scenarios. (v) Two objective model evaluation experiments are conducted on the two databases, respectively, and experimental results demonstrate the effectiveness of our proposed methods.

Our data collection softwares, two databases, benchmark models, as well as objective metrics will be released to facilitate future research. We hope this study will motivate other researchers to consider both the see-through view and the augmented view when conducting AR QoE studies.

II. Related Work and Organizations of This Paper

A. Augmented Reality and Visual Confusion

This work mainly concerns head-mounted AR applications rather than mobile phone based AR applications. Considering rendering methods, there are two main aspects of works in the field of AR visualization, including 2D displaying and 3D rendering. The superimposition between real-world scenes and virtual contents may cause visual confusion. Depending on display methods (*i.e.*, superimposition methods), AR technology may produce two kinds of visual effects, including binocular visual confusion and monocular visual confusion. We discuss the relationship between these aspects and our work as follows.

2D displaying. The most basic application of AR is displaying digital contents in a 2D virtual plane [13]. These digital contents include images, videos, texts, shapes, and even 3D objects in 2D format, *etc.* To display 2D digital contents, a real world plane is needed to attach the virtual plane. The real world plane and virtual plane are usually in one same *Vieth–Müller* circle (*a.k.a,* isovergence circle), which may cause visual confusion. This situation is the main consideration of this paper.

Fig. 2. Relationship between CFIQA and ARIQA.

3D rendering. Compared to 2D displaying, 3D rendering aims at providing 3D depth cues of virtual objects (note that this depth cue is a little bit different from the depth of the above mentioned virtual plane) [14]. Although the 3D depth cues will cause the real world scenes and virtual objects to be located in different *Vieth–Müller* circles, the visual confusion effect still exists, which makes this situation more complex (see Section VII for more details).

Visual Confusion. Visual confusion is the perception of two different images superimposed onto the same space [15]. The concept of visual confusion comes from ophthalmology, which is usually used to describe the perception of diplopia (*i.e.*, double vision) caused by strabismus [16]. Note that in this paper, we only consider the "seeing of two or more different views/things in one direction" [11], [12] as the definition of visual confusion, which should be distinguished from visual illusions [17] and most perceptual distortions. However, some distortions such as ghosting artifacts can be regraded as visual confusion. Although the multiplexing superimposition can extend visual perception and has been widely used in prism designs for field expansion [12], some studies have reported that the visual confusion that occurs during field expansion makes users uncomfortable, annoying and disturbing [11], [18]. Thus, it is significant to study the visual confusion effect for AR QoE assessment.

Binocular visual confusion. The visual confusion caused by two views superimposed binocularly (within two eyes respectively) is *binocular* visual confusion, which may lead to binocular rivalry [12], [19]. Some previous monocular AR devices, such as Google Glass [20] and VUZIX M400 [21], were mainly constructed based on binocular visual superimposition to avoid the occlusion produced by AR devices. However, the binocular rivalry caused by binocular visual confusion may strongly affect the QoE [22].

Monocular visual confusion. The visual confusion caused by two views superimposed monocularly (within one eye) is *monocular* visual confusion [12], which may lead to monocular rivalry [12], [19]. Since monocular rivalry is much weaker than binocular rivalry [22] and it possibly occurs only with extended attention [12], most recent binocular AR technologies were built based on monocular visual superimposition to avoid occlusion, such as Microsoft HoloLens [23], Magic Leap [24], Epson AR [25], *etc.* However, the QoE of monocular visual confusion still lacks thorough discussion, which is mainly considered in this paper.

B. Image Quality Assessment

Many IQA methods have been proposed in the past decades [26], [27], which can be roughly grouped into three categories, including full reference (FR) IQA, reduced reference (RR) IQA, and no reference (NR) IQA. Considering the possible application scenarios where the digital contents and real-world scenes can be easily obtained, in this paper, we mainly focus on the FR-IQA metric.

Classical IQA index. In terms of FR IQA methods, many classical metrics have been proposed and widely used, including mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) index [28], feature similarity (FSIM) index [29], *etc.* Regarding NR IQA indexes, there are also many popular methods such as natural image quality evaluator (NIQE) [30], blind quality assessment based on pseudo-reference image (BPRI) [31], and NR free-energy based robust metric (NFERM) [32], *etc.*

Learnable IQA. Driven by the rapid development of deep neural networks (DNNs) recently, some learning based IQA algorithms have been proposed. Kang *et al.* [33] proposed to use multi-task convolutional neural networks to evaluate the image quality and use 32×32 patches for training. Bosse *et al.* [34] proposed both FR and NR IQA metrics by joint learning of local quality and local weights. Some studies located specific distortions by using convolutional sparse coding and then evaluated the image quality [35]. Recently, some studies also demonstrated the effectiveness of using pre-trained DNN features in calculating visual similarity [36], [37].

AR/VR IQA. As discussed in the introduction, most previous AR/VR IQA works have studied the degradations of geometry and texture of 3D meshes and point clouds inside AR/VR [8]–[10]. Unlike AR research, many works on VR IQA have also investigated the quality of omnidirectional images (*a.k.a, equirectangular images*) [38]–[41], since the format of these images is different with traditional images. In this paper, we propose that in AR technology, confusing image is its special "image format", and confusing image quality assessment is equally important with 3D meshes or point clouds quality assessment, since it is not only related to 2D displaying but also associated with 3D rendering effects. Moreover, it significantly influences the QoE of AR.

C. Relationship Between CFIQA and ARIQA

Fig. 2 illustrates the relationship between CFIQA and ARIQA. CFIQA is a more basic and general problem, which aims at predicting the image quality for each image layer of the superimposition between any two images. Note that superimposed images/scenes are frequently encountered in the real-world, *e.g.*, rain, haze, reflection, *etc*, [42]–[44], but we mainly focus on the superimposition between two complex images in this paper, since it is directly related to AR applications. ARIQA is a more specific application scenario of CFIQA, which mainly considers the image quality of the AR layer, since we are more concerned with the saliency or visibility problem of the BG layer [13], [45]. This saliency or visibility issue is not something we discuss and address in this paper (see Section VII for more details). Moreover, in this paper, the FOV of two image layers in CFIQA is the same since it is a more general problem and the influence of the superimposition on perceptual quality is unknown, while the FOV of BG scenes is larger than that of AR contents in ARIQA since it is a more realistic situation in AR display [2], [45].

The rest of the paper is organized as follows. Section III describes the construction procedure of the CFIQA database.

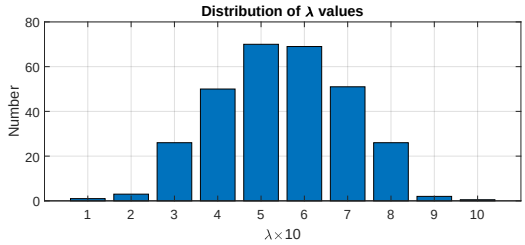


Fig. 3. Distribution of λ values in this paper. The values are multiplied by 10 and rounded up.

Fig. 4. Illustration of the subjective experiment interface for CFIQA.

Section IV introduces the objective CFIQA models including benchmark models and the proposed CFIQA model. The experimental validation procedure and results of these objective CFIQA models are given in Section V. Then an extended subjective & objective ARIQA study is presented in Section VI. Section VII concludes the paper and discusses several future issues.

III. Subjective ConFusing Image Quality Assessment (CFIQA)

A. Confusing Image Collection

To address the problem of subjective confusing IQA data absence, we first build a novel ConFusing Image Quality Assessment (CFIQA) database. Since this paper is the first work to study confusing IQA, we consider the visual confusion as the only distortion type in this section to study whether and how visual confusion influences the perceptual quality. We collect 600 images from Pascal VOC dataset [46] as reference images and split them into two groups. Then we randomly select two reference images from these two groups and mixed them in pair with a blending parameter λ to generate a distorted image. This can be formulated as:

$$I_D = \lambda \circ I_{R_1} + (1-\lambda) \circ I_{R_2} \ , \tag{1}$$

where I_{R_1} is the reference image from the first group, I_{R_2} is the reference image from the second group, $\lambda \in [0,1]$ represents the degradation value of mixing, I_D denotes the generated distorted image. All reference images are resized to the size of 512×512 and then superimposed. A total of 300 distorted images are finally generated.

Obviously, λ value near 0 or 1 will cause one image to be unnoticeable while closer to the center (*i.e.,* 0.5) will cause near confusion for both views. Since visual confusion is the main consideration in this section, it is unreasonable to randomly sample λ from $[0,1]$. In this work, to make the λ value to be closer to the center values in range $[0,1]$, we sampled λ value from a Beta distribution, *i.e.,* $\lambda \sim \mathbf{Beta}(\alpha,\alpha)$. The parameter α is set to 5 in this database. Fig. 3 demonstrates the distribution of λ values used in the CFIQA database.

B. Subjective Experiment Methodology

Experiment setup. A subjective experiment is conducted on the dataset. There are several subjective assessment methodologies recommended by the ITU [47], for instance, single-stimulus (SS), double-stimulus impairment scale (DSIS) and paired comparison (PC). Since the reference images are available, we adopt a paired comparison continuous quality evaluation (PCCQE) strategy to obtain the subjective quality ratings. As shown in Fig. 4, for each distorted image, we display its two reference images simultaneously and instruct the subjects to give two opinion scores of the perceptual quality of two layer views (*i.e.,* two reference images) in the distorted image, respectively. We suggest the subjects to only view the distorted image in the center, and two reference images are just used to determine the quality of which layer is being given. Two continuous quality rating bars are presented to the subject. The quality bar is labeled with five Likert adjectives: Bad, Poor, Fair, Good and Excellent, allowing subjects to smoothly drag a slider (initially centered) along the continuous quality bar to select their ratings. They are seated at a distance of about 2 feet from the monitor, and this viewing distance is roughly maintained during each session. All images are shown in their raw sizes, *i.e.,* 512×512 with random sequence during the experiment.

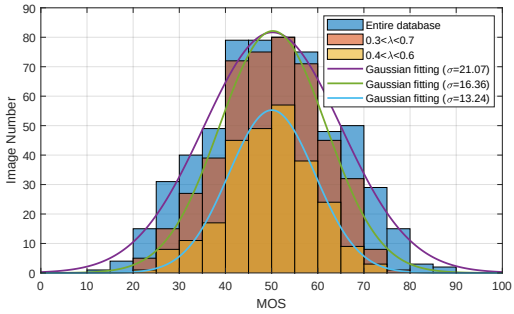


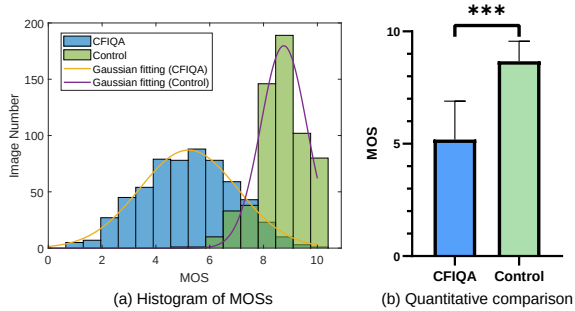
Fig. 5. Histogram of MOSs from the CFIQA database within different λ value ranges.

Testing procedure. As suggested by ITU [47], at least 15 subjects are required to conduct subjective IQA experiment. A total of 17 subjects are recruited to participate in the study. Before participating in the test, each subject have read and signed a consent form which explained the human study. All subjects are determined to have normal or corrected-to-normal vision. General information about the study is supplied in printed form to the subjects, along with instructions on how to participate in the task. Each subject then experiences a short training session where 20 confusing images (not included in the formal test) are shown, allowing them to become familiar with the user interface and the visual confusion distortion which may occur. Moreover, subjects have enough rest time every 10 minutes to avoid fatigue during the experiment.

C. Subjective Data Processing and Analysis

We follow the suggestions given in [47] to conduct the outlier detection and subject rejection. Specifically, we first calculate the kurtosis score of the raw subjective quality ratings for each image to detect it is a Gaussian case or a non-Gaussian case. Then, for the Gaussian case, the raw score for an image is considered to be an outlier if it is outside 2 standard deviations (stds) about the mean score of that image; for the non-Gaussian case, it is regarded as an outlier if it is outside $\sqrt{20}$ stds about the mean score of that image. A subject

Fig. 6. Sample images from CFIQA database. MOS, SSIM, FSIM values, as well as λ value are given in the figure. Note that a MOS in this figure mean the MOS of the reference layer in the distorted image.



(a) Histogram of MOSs (b) Quantitative comparison

Fig. 7. Comparisons of MOSs between the subjective CFIQA experiment and the controlled experiment.

is removed if more than 5% of his/her evaluations are outliers. As a result, only 1 subject is rejected, and each image is rated by 16 valid subjects. Among all scores given by the remaining valid subjects, about 2.77% of the total subjective evaluations are identified as outliers and are subsequently removed. For the remaining 16 valid subjects, we convert the raw ratings into Z-scores, which are then linearly scaled to the range $[0, 100]$ and averaged over subjects to obtain the final mean opinion scores (MOSs) as follows:

$$z_{ij} = \frac{m_{ij} - \mu_i}{\sigma_i}, \quad z'_{ij} = \frac{100(z_{ij} + 3)}{6}, \tag{2}$$

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} z'_{ij}, \tag{3}$$

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

Fig. 5 plots the histograms of MOSs over the entire database as well as within different λ value ranges, showing a wide range of perceptual quality scores. We also plot the Gaussian curve fitting for the histogram. It can be observed that when λ is closer to 0.5, the MOSs tend to be more centered (illustrated by the smaller σ value), but still have a wide range of perceptual quality scores.

Based on the constructed database, we further qualitatively analyze the characteristic of the visual confusion. As shown in Fig. 6, we roughly classify the visual confusion into three categories. The first category is "strong confusion", which means that the mixing of two reference layers will cause strong confusion and may affect the quality rating of the superimposed image (distorted image). The second category is "confusion but acceptable", which represents that the visual confusion caused by the superimposition of two reference layers is acceptable, or uninfluenced, or the perceptual quality is even improved. The third category is "suppression", which denotes that in the superimposed image, one reference layer will suppress another reference layer. This results in the situation that the perceptual quality of one layer is much better than another layer. First of all, as a general observation from Fig. 6, we notice that the MOS values (*i.e.,* subjective quality scores) are commonly sorted in descending order as: (1) the activated layer in the "suppression" category (*i.e.,* the clearer layer), (2) two image layers in the "confusion but acceptable" category, (3) two image layers in the "strong confusion" category, and (4) the suppressed layer in the "suppression" category (*i.e.,* the fainter layer). Furthermore, with thorough observation, we notice that the saliency relation between two layers of the superimposed image are important to the perceptual quality, which are demonstrated by several IQA metrics in Section V. From the above observations, we conclude that without introducing other distortions, visual confusion itself can significantly influence the quality of confusing images.

D. Controlled Experiment

To further validate the assumption that the superimposition between two images will cause visual confusion and significantly influence the perceptual quality, we also conduct a controlled experiment based on pristine images and compare the MOS results with those of the above subjective CFIQA experiment. Specifically, a new group of 17 subjects is recruited for the controlled experiments. These subjects are asked to give their opinion scores for all pristine images of the superimposed images used in the previous subjective CFIQA experiment. Then the MOSs of these pristine images are calculated correspondingly. Fig. 7 demonstrates the comparison results between this controlled experiment and the above CFIQA experiment. As shown in Fig. 7 (a), the MOS distribution of the controlled experiment is more centered on the right side of this figure compared with that of the CFIQA experiment. Fig. 7 (b) further demonstrates the paired t-test comparison result between the MOSs of the CFIQA experiment and those of the controlled experiment, which illustrates that the MOSs of the controlled experiment are significantly larger ($p < 0.0001$). These experimental

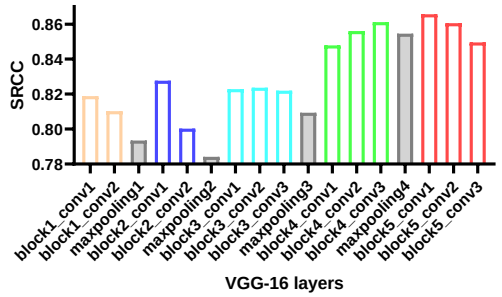


Fig. 8. The SRCC scores between the distance scores and the subjective ratings for each layer l of the VGG-16 network.

results further quantitatively validate our assumption that the visual confusion caused by the superimposition significantly influences the perceptual quality.

IV. Objective CFIQA Models

A. Benchmark Models

In terms of the CFIQA database, one distorted superimposed image corresponds to two reference image layers and two subjective ratings. Thus, we modify state-of-the-art FR-IQA metrics as benchmark models to cope with the CFIQA database as follows.

The simplest metric. The first intuitive idea is applying the mixing value λ as the metric for evaluating visual confusion since it is directly related to the generation of distorted images. As shown in Eq. (1), larger λ values make image I_D and image I_{R_1} more similar, and make image I_D and image I_{R_2} more different, while larger $(1-\lambda)$ values make image I_D and image I_{R_2} more similar, and make image I_D and image I_{R_1} more different. Therefore, for a distorted image I_D generated by two image layers I_{R_1} and I_{R_2} via Eq. (1), the simplest FR metric is using λ and $(1-\lambda)$ to predict the perceptual qualities of I_{R_1} and I_{R_2} in I_D, respectively.

Classical FR-IQA metrics. We test 16 state-of-the-art classical FR-IQA metrics on the CFIQA database, including MSE, PSNR, NQM [48], SSIM [28], IFC [49], VIF [50], IW-MSE [51], IW-PSNR [51], IW-SSIM [51], FSIM [29], GSI [52], GMSD [53], GMSM [53], PAMSE [54], LTG [55], and VSI [56]. Considering that for a distorted image, there are two corresponding reference images, we calculate the FR similarities of this distorted image and the two reference images respectively to obtain two predicted quality scores. Moreover, since visual attention is important in this task as discussed above, we further select 3 widely used and well-performed metrics, *i.e.*, SSIM, FSIM, and GMSM, and incorporate saliency weights into the quality pooling as new metrics, which are denoted as "SSIM + saliency", "FSIM + saliency", and "GMSM + saliency", respectively.

Deep feature based IQA metrics. Recently, many works demonstrate the consistence between DNNs and human visual perception [36], [37], [57]. Therefore, we also consider modifying these DNNs as benchmark models for objective CFIQA. We first build baseline models with several state-of-the-art DNNs, including SqueezeNet [58], AlexNet [59], VGG (VGG-16 and VGG-19) [60], and ResNet (ResNet-18, ResNet-34, and ResNet-50) [61], which are denoted as "Baseline (SqueezeNet)", "Baseline (AlexNet)", "Baseline (VGG-16)", *etc*. These baseline models are constructed by averaging the subtracted features of the selected layers (see Section IV-B, Para. 2 for more details) between distorted images and reference images as follows:

$$d(l) = \frac{1}{H_l W_l C_l} \sum_{h,w,c} f_d^l, \tag{4}$$

where f_d^l is the subtracted feature vector for the selected l-th layer (see Eq. (5) in Section IV-B, Para. 3 for more details), and H_l, W_l, C_l are the height, width, and number of channels of the l-th layer. Considering the features extracted from the last layer of each "component" of these networks may not well reflect the overall performance, in this work, we propose a method to improve the baseline performance based on VGG-16, which is denoted as "Baseline+ (VGG-16)". As illustrated in Fig. 8, for each layer l of the 17 layers (13 convolutional layers and 4 max pooling layers) of VGG-16, we calculate the Spearman Rank-order Correlation Coefficient (SRCC) between $d(l)$ and the MOSs over the whole dataset, which allows us to observe if a particular layer of the model provides more relevant feature maps for CFIQA. Then the 5 most correlated layers are extracted and averaged to compute the predicted score of the model "Baseline+ (VGG-16)". Furthermore, two widely used deep feature fusion-based metrics (*i.e.*, LPIPS [37] and DISTS [62]) are also included as benchmark models.

Overall, for each benchmark model, we obtain 600 predicted quality scores corresponding to 600 subjective quality ratings (MOSs) for 300 distorted images. The performance of each model is then calculated between all these predicted quality scores and subjective quality ratings using the criteria discussed in Section V.

B. Attention Based Deep Feature Fusion Method (The Proposed CFIQA Model)

As shown in Fig. 6, the classical SSIM index and FSIM index are inconsistent with human perception in some cases (see Section V for more quantitative comparison). From the above analysis, we suppose that the assessment of visual confusion is related to both low-level visual characteristics and high-level visual semantic features, since the visual confusion may disturb semantic information and affect the perceptual quality. This may cause the failure of the classical metrics, since most of them only consider low-level visual features. As discussed in [36], [37], deep features also demonstrate great effectiveness as perceptual metrics. Moreover, DNN can extract both low-level and high-level features. Therefore, in this paper, we propose an attention based deep feature fusion method to measure the visual confusion, which is shown in Fig. 9.

Deep feature extraction. We first employ several state-of-the-art pre-trained DNNs to extract both low-level and high-level features, which include SqueezeNet [58], AlexNet [59], VGG Net [60], and ResNet [61]. SqueezeNet is an extremely lightweight DNN with comparable performance on classification benchmark. The features from the first $conv$ layer and subsequent "$fire$" modules in SqueezeNet are extracted and used. We also use a shallow AlexNet network which may more closely match the architecture of

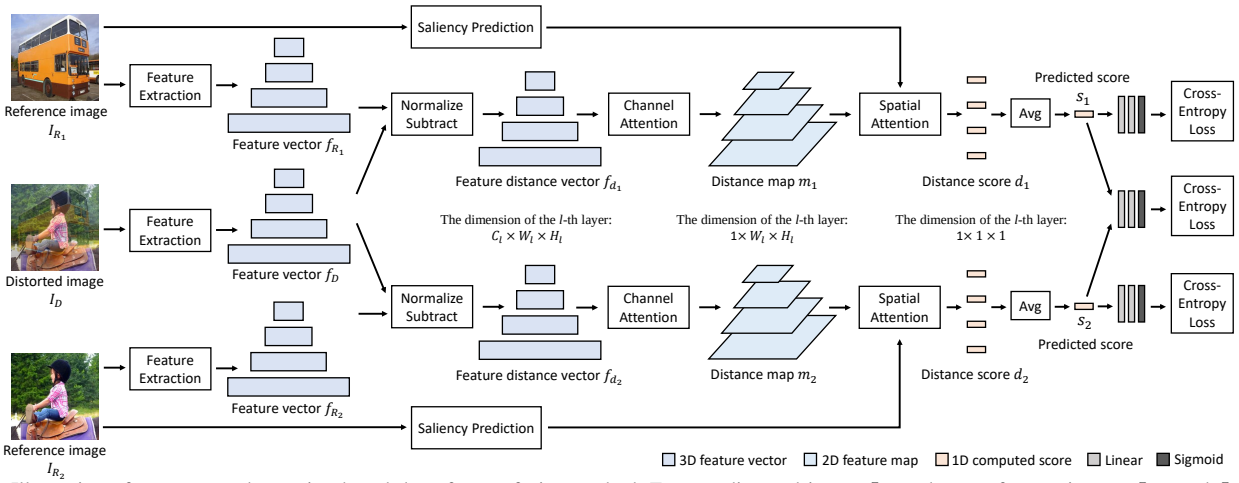


Fig. 9. Illustration of our proposed attention based deep feature fusion method. For one distorted image I_D and two reference images I_{R_1} and I_{R_2}, three DNN feature vectors f_D, f_{R_1}, and f_{R_2} are first extracted. We then compute the feature distances between the corresponding feature layers of f_D and f_{R_1}, as well as f_D and f_{R_2}, respectively, to get distance vectors f_{d_1}, f_{d_2}. Next, two feature distance vectors f_{d_1}, f_{d_2} are fed into a channel attention module to get distance map stacks m_1, m_2. After weighting by a spatial attention operation, two score vectors d_1 and d_2 are computed. Finally, two predicted scores s_1 and s_2 are calculated by averaging d_1 and d_2 respectively. Two kinds of loss functions are used to constrain the learning process, including the ranking loss (middle), and the score regression loss (top and bottom). All loss functions are based on the cross-entropy loss.

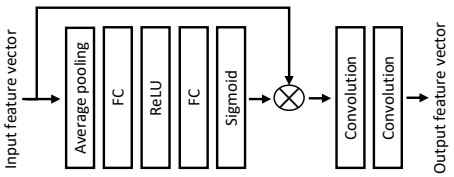


Fig. 10. Illustration of the channel attention module used in the proposed CFIQA model.

the human visual cortex [63], and we extract the features from the $conv1 - conv5$ layers in AlexNet. Furthermore, 5 *conv* layers labeled $conv1\_2$, $conv2\_2$, $conv3\_3/conv3\_4$, $conv4\_3/conv4\_4$, $conv5\_3/conv5\_4$ are extracted from two VGG networks (VGG-16 and VGG-19), respectively. Finally, considering the effectiveness of ResNet in a large amount of computer-vision (CV) tasks, we also explore the utility of features extracted by ResNet in this task. Three ResNet architectures including ResNet-18, ResNet-34, as well as ResNet-50 are considered. We use the features extracted from the first $conv$ layer and subsequent "$BasicBlock$" or "$Bottleneck$" modules for all of these architectures. Overall, 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.

Computing feature distance. Then we follow the method in [37] and calculate the feature distance vectors between a distorted image and two reference images by subtracting normalized feature stacks. This can be expressed as:

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

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 vector.

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 [64] method to learn and re-organize features. As shown in Fig. 10, after the channel attention, two convolutional layers are followed to reduce channel numbers. The kernel size of all convolutional layers is set to 1. Through this manipulation, a distance map stack m_i can be obtained for each f_{d_i}, where $i \in \{1, 2\}$.

Spatial attention. As discussed in Section III-C, 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 [65] is used to calculate the spatial attention map W_i for a reference image I_{R_i}. By weighting the distance map m_i with a scaled spatial attention map W_i, the distance score for each layer l can be calculated as:

$$d_i^l = \frac{\sum_{h,w} W_{i\,hw}^l \odot m_{i\,hw}^l}{\sum_{h,w} W_{i\,hw}^l} \ . \tag{6}$$

Then the final quality score can be computed as:

$$s_i = \mathtt{Avg}_l(d_i^l) \ , \tag{7}$$

where $\mathtt{Avg}$ is the average operation, and s_i is the predicted quality score.

Loss function. Different with [37], which aims at two alternative forced choice (2AFC) test, our work focuses on a more general quality assessment task, *i.e.,* the MOS prediction task. Moreover, different from traditional IQA condition, in our dataset, one distorted image corresponds to two reference images. Thus, the loss function needs to be carefully designed. An intuitive loss function is to compare (rank) the perceptual qualities of two layers in the distorted image. Therefore, a ***ranking loss*** $\mathcal{L}_R$ is adopted to predict the probability that one layer suppress another layer, which is based on the cross-entropy loss function. Although the above ranking loss can predict the relative perceptual quality of two layers in a distorted image, the overall qualities of different images across the whole dataset are not normalized and compared. Therefore, another ***score regression loss*** $\mathcal{L}_S$ is introduced to regress the probability of the quality being bad or excellent, which is also built based on the cross-entropy loss function. Two linear

TABLE I
PERFORMANCE COMPARISON OF THE STATE-OF-THE-ART FR-IQA MODELS ON THE CFIQA DATASET. THE DATASET IS SPLIT INTO THREE FOLDS, INCLUDING "ENTIRE", "$0.3< \lambda <0.7$", AND "$0.4< \lambda <0.6$". WE **BOLD** THE BEST TWO RESULTS IN EACH GROUP OF MODELS. "*" MEANS A RE-TRAINED MODEL.

Dataset	Entire					$0.3< \lambda <0.7$					$0.4< \lambda <0.6$				
Model \ Criteria	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑
λ	0.7700	0.5881	0.8021	8.2909	11.130	0.5976	0.4283	0.6097	8.5426	7.6935	0.3597	0.2538	0.3698	8.6638	4.4961
MSE	0.6944	0.5072	0.7215	9.6118	9.8519	0.5030	0.3448	0.5073	9.3010	6.1129	0.2816	0.1909	0.3555	8.7250	3.3331
PSNR	0.6944	0.5072	0.7196	9.6399	9.8348	0.5030	0.3448	0.5076	9.2988	6.1341	0.2816	0.1909	0.3053	8.8574	3.3129
NQM [48]	0.7486	0.5613	0.7683	8.8752	10.613	0.6012	0.4272	0.6131	8.5048	7.4643	0.4742	0.3282	0.5150	7.9967	5.4664
SSIM [28]	0.7907	0.6005	0.8179	7.9970	11.204	0.6520	0.4660	0.6707	8.0269	8.0014	0.4858	0.3394	0.5143	7.9994	5.1510
IFC [49]	0.7198	0.5362	0.7439	9.2430	10.461	0.5971	0.4253	0.6118	8.5491	7.4499	0.5416	0.3881	0.5610	7.6933	5.8762
VIF [50]	**0.8346**	**0.6491**	**0.8556**	**7.1762**	**11.811**	**0.7266**	**0.5349**	**0.7372**	**7.3064**	**8.8275**	0.6244	**0.4548**	**0.6505**	**7.0875**	6.5839
IW-MSE [51]	0.7468	0.5563	0.7777	8.7497	10.592	0.5787	0.4019	0.5999	8.6496	7.0436	0.3524	0.2348	0.3865	8.6132	3.6707
IW-PSNR [51]	0.7468	0.5563	0.7694	8.8837	10.571	0.5787	0.4019	0.6008	8.6407	7.0474	0.3524	0.2348	0.3871	8.6105	3.5713
IW-SSIM [51]	**0.8519**	**0.6715**	**0.8700**	**6.8096**	**12.015**	**0.7512**	**0.5607**	**0.7638**	**6.9567**	**9.0641**	**0.6633**	**0.4804**	**0.6856**	**6.7796**	**7.0594**
FSIM [29]	0.8331	0.6486	0.8540	7.2151	11.721	0.7201	0.5268	0.7263	7.4205	8.6681	**0.6296**	0.4501	0.6496	7.0917	**6.7445**
GSI [52]	0.7904	0.6002	0.8152	8.0506	11.147	0.6507	0.4623	0.6656	8.0538	7.8943	0.5085	0.3516	0.5447	7.8056	5.5417
GMSD [53]	0.7609	0.5725	0.7894	8.5356	10.812	0.6343	0.4533	0.6472	8.2345	7.7938	0.5678	0.3986	0.6152	7.3588	6.2830
GMSM [53]	0.8117	0.6238	0.8328	7.6853	11.433	0.6930	0.5013	0.7022	7.6849	8.3621	0.5866	0.4122	0.6070	7.4067	6.3127
PAMSE [54]	0.6827	0.4965	0.7104	9.7694	9.6744	0.4855	0.3323	0.4930	9.3900	5.8931	0.2648	0.1802	0.3306	8.8155	3.0792
LTG [55]	0.7922	0.6049	0.8160	8.0271	11.189	0.6672	0.4806	0.6779	7.9397	8.1256	0.5731	0.4071	0.6121	7.3762	6.3452
VSI [56]	0.8297	0.6455	0.8544	7.2150	11.679	0.7106	0.5188	0.7256	7.4266	8.6095	0.5860	0.4175	0.6285	7.2427	6.3984
SSIM [28] + saliency	**0.8364**	**0.6520**	**0.8528**	**7.2437**	**11.739**	**0.7228**	**0.5318**	**0.7362**	**7.3157**	**8.6880**	0.5797	0.4115	0.6063	7.4211	6.0386
FSIM [29] + saliency	**0.8542**	**0.6705**	**0.8711**	**6.8162**	**11.926**	**0.7558**	**0.5584**	**0.7664**	**6.9428**	**8.9966**	**0.6628**	**0.4724**	**0.6728**	**6.9110**	**6.9979**
GMSM [53] + saliency	0.8339	0.6470	0.8519	7.2608	11.672	0.7227	0.5268	0.7336	7.3292	8.5930	**0.6176**	**0.4354**	**0.6223**	**7.2978**	**6.4243**
LPIPS (Squeeze) [37]	0.8338	0.6468	0.8521	7.2596	11.676	0.7278	0.5341	0.7418	7.2346	8.7265	0.6117	0.4399	0.6708	6.9190	6.5814
LPIPS (Alex) [37]	0.8474	0.6610	0.8612	7.0603	11.847	0.7546	0.5615	0.7647	6.9547	9.0461	0.6583	0.4793	0.6943	6.7094	7.2792
LPIPS (VGG) [37]	0.8376	0.6512	0.8520	7.2607	11.762	0.7439	0.5501	0.7535	7.0982	8.9189	0.6551	0.4772	0.7013	6.6587	6.9437
LPIPS (VGG) [37] *	0.8413	0.6549	0.8607	7.0817	11.799	0.7508	0.5550	0.7596	7.0379	8.9663	0.6673	0.4881	0.6950	6.7167	7.0033
DISTS [62]	0.7709	0.5815	0.7967	8.4070	10.684	0.6526	0.4658	0.6810	7.9190	7.7870	0.5253	0.3665	0.5531	7.7848	5.7042
Baseline (SqueezeNet)	0.8625	0.6806	0.8774	6.6574	12.038	0.7760	0.5823	0.7822	6.7219	9.2077	0.6902	0.5046	0.7170	6.5044	7.4181
Baseline (AlexNet)	0.8752	0.6987	0.8870	6.4044	12.189	0.7967	0.6061	0.8049	6.4086	9.4289	0.7162	0.5321	0.7364	6.3125	7.4786
Baseline (VGG-16)	0.8494	0.6655	0.8663	6.9274	11.944	0.7630	0.5683	0.7712	6.8757	9.1520	0.6792	0.4970	0.7108	6.5686	7.4597
Baseline+ (VGG-16)	0.8629	0.6820	0.8772	6.6615	12.078	0.7847	0.5910	0.7921	6.5973	9.3582	0.7028	0.5167	0.7265	6.4195	7.4188
Baseline (VGG-19)	0.8622	0.6798	0.8769	6.6573	12.106	0.7804	0.5847	0.7859	6.6754	9.3364	0.7010	0.5126	0.7327	6.3490	7.4755
Baseline (ResNet-18)	**0.8784**	**0.7004**	**0.8920**	**6.2736**	**12.236**	**0.8025**	**0.6079**	**0.8087**	**6.3598**	**9.5158**	0.7227	0.5318	0.7393	6.2907	7.4336
Baseline (ResNet-34)	**0.8823**	**0.7053**	**0.8941**	**6.2185**	**12.273**	**0.8098**	**0.6157**	**0.8137**	**6.2847**	**9.5477**	**0.7405**	**0.5490**	**0.7570**	**6.1058**	**7.6783**
Baseline (ResNet-50)	0.8759	0.6963	0.8885	6.3699	12.215	0.8011	0.6060	0.8060	6.3992	9.4857	**0.7250**	**0.5340**	**0.7447**	**6.2347**	**7.5612**
CFIQA (SqueezeNet)	0.8712	0.6933	0.8827	6.5180	12.112	0.7847	0.5910	0.7933	6.5641	9.2214	0.6904	0.5050	0.6994	6.6646	6.9383
CFIQA (AlexNet)	0.9053	0.7347	0.9109	5.7467	12.466	0.8438	0.6520	0.8461	5.7758	9.7944	0.7749	0.5833	0.7878	5.7559	7.8676
CFIQA (VGG-16)	0.9203	0.7550	**0.9258**	**5.2636**	**12.652**	0.8681	0.6782	0.8700	5.3412	10.012	0.8055	0.6105	0.8168	5.3859	8.1268
CFIQA (VGG-19)	0.9178	0.7527	0.9229	5.3597	12.611	0.8643	0.6762	0.8646	5.4443	9.9785	0.7987	0.6068	0.8119	5.4520	8.0631
CFIQA (ResNet-18)	0.9196	0.7580	0.9256	5.2713	12.632	0.8677	0.6844	0.8688	5.3661	9.9947	0.8118	0.6272	0.8291	5.2225	8.1516
CFIQA (ResNet-34)	**0.9209**	**0.7604**	0.9256	5.2746	12.645	**0.8703**	**0.6883**	**0.8706**	**5.3293**	**10.032**	**0.8204**	**0.6381**	**0.8310**	**5.1876**	**8.2088**
CFIQA+ (ResNet-34)	**0.9232**	**0.7625**	**0.9278**	**5.1966**	**12.668**	**0.8733**	**0.6897**	**0.8741**	**5.2624**	**10.058**	**0.8226**	**0.6341**	**0.8326**	**5.1592**	**8.2436**
CFIQA (ResNet-50)	0.8964	0.7241	0.9042	5.9441	12.402	0.8275	0.6370	0.8311	6.0228	9.6863	0.7664	0.5774	0.7929	5.6932	7.9248

layers and a sigmoid layer are used to project the predicted value to the ground-truth space. The overall loss function can be formulated as:

$$\mathcal{L} = \mathcal{L}_{S_1} + \mathcal{L}_{S_2} + \gamma \mathcal{L}_R, \tag{8}$$

where the hyperparameter γ is empirically set as 2.

Edge feature integration. The edges of objects can help identify their categories [66]. 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 [66] and concatenate them with the features extracted from one of the aforementioned classification backbones as an enhanced model, which is named CFIQA+.

V. EXPERIMENTAL VALIDATION ON THE CFIQA DATABASE

A. Experimental Protocol

Experimental settings. As mentioned in Section IV-B, our proposed method needs some samples to train the feature fusion module. Therefore, the CFIQA database is split into two parts at a ratio of 1:1, *i.e.*, each part has 150 distorted images and corresponding 300 reference images. A two-fold cross-validation experiment is conducted. The two splits of the database are used as the training set and the test set respectively in each cross-validation fold. During training, the feature extraction network is frozen, and we only train the feature fusion part. We use Adam as the optimizer [67]. The network is trained for 100 epoch with a learning rate of 0.0001, and additional 50 epoch with decayed learning rate from 0.0001 to 0. The batch size is set as 10 during training.

Traditional evaluation metrics. To evaluate the various quality predictors, we use a five-parameter logistic function to fit the quality scores:

$$Q' = \beta_1 (\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (Q - \beta_3)}}) + \beta_4 Q + \beta_5, \tag{9}$$

where Q and Q' are the objective and best-fitting quality, $\beta_i (i = 1, 2, 3, 4, 5)$ are the parameters to be fitted during the evaluation. Five evaluation metrics are adopted to measure the consistency between the ground-truth subjective ratings and the fitted quality scores, including Spearman Rank-order Correlation Coefficient (SRCC), Kendall Rank-order Correlation Coefficient (KRCC), Pearson Linear Correlation Coefficient (PLCC), Root Mean Square Error (RMSE) and Perceptually Weighted Rank Correlation (PWRC) [68]. The evaluation is conducted on the entire database and 2 sub-datasets within different λ ranges. Note that evaluating the performance in the range that the λ value is near 0.5 is significant, since it is near the essence of visual confusion and is more common

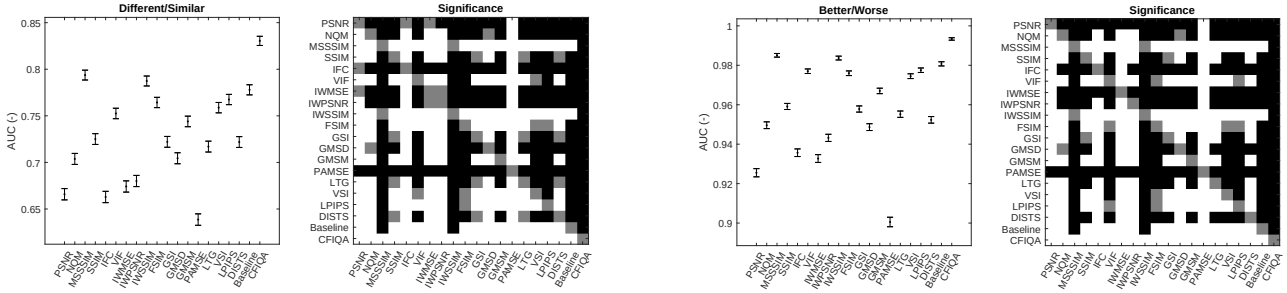


Fig. 11. New criteria performance of 19 state-of-art FR IQA models and the proposed metric on the CFIQA database. Left two figures are the different vs. similar ROC analysis results. Right two figures are the better vs. worse analysis results. Note that a white/black square in the significance figures means the row metric is statistically better/worse than the column one. A gray square means the row method and the column method are statistically indistinguishable. The backbone of all networks in these figures is VGG-16.

TABLE II
IMPACT OF DIFFERENT COMPONENTS. (BACKBONE: VGG-16. "CA" MEANS "CHANNEL ATTENTION". "SA" MEANS "SPATIAL ATTENTION").

Dataset	Entire					$0.3< \lambda <0.7$					$0.4< \lambda <0.6$				
Model \ Criteria	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑
only ranking loss	0.8413	0.6549	0.8607	7.0817	11.799	0.7508	0.5550	0.7596	7.0379	8.9663	0.6673	0.4881	0.6950	6.7167	7.0033
w/o CA & SA	0.8647	0.6830	0.8794	6.5994	12.103	0.7834	0.5875	0.7885	6.6417	9.3234	0.6973	0.5122	0.7306	6.3735	7.3048
w/o SA	0.8797	0.7015	0.8917	6.2746	12.270	0.8046	0.6094	0.8094	6.3477	9.5354	0.7216	0.5326	0.7458	6.2209	7.4950
w/o CA	0.9160	0.7490	0.9216	5.4075	12.597	0.8618	0.6711	0.8612	5.5065	9.9569	0.7925	0.5982	0.8039	5.5480	8.0077
w/o ranking loss	0.9069	0.7363	0.9139	5.6435	12.507	0.8473	0.6545	0.8469	5.7586	9.8297	0.7715	0.5746	0.7864	5.7731	7.8845
w/o scoring loss	0.9139	0.7466	0.9195	5.4042	12.611	0.8600	0.6697	0.8618	5.4442	9.9765	0.7917	0.5982	0.8076	5.4694	8.0395
all combined	**0.9203**	**0.7550**	**0.9258**	**5.2636**	**12.652**	**0.8681**	**0.6782**	**0.8700**	**5.3412**	**10.011**	**0.8055**	**0.6105**	**0.8168**	**5.3859**	**8.1268**

in AR applications, while the poor cases in this situation may strongly influence the QoE. However, most FR-IQA metrics cannot perform well in this range, which may limit their practicality on CFIQA.

New evaluation methodology. As a complementary, the receiver operating characteristic (ROC) analysis methodology [69], [70] is also adopted for metric evaluation, which is based on two aspects, *i.e.*, *whether two stimuli are qualitatively different and if they are, which of them is of higher quality*. The Fig. 1. in the *supplementary material* illustrates the framework of this evaluation methodology. We first conduct pair-wise comparison for all possible image pairs, and then classify them into pairs with and without significant quality differences. Then the ROC analysis is used to determine whether various objective metrics can discriminate images with and without significant differences, termed "*Different vs. Similar ROC Analysis*". Next, the image pairs with significant differences are classified into pairs with positive and negative differences, and the ROC analysis is used to test if various objective metrics can distinguish images with positive and negative differences, termed "*Better vs. Worse ROC Analysis*". The area under the ROC curve (AUC) values of two analysis are mainly reported in this paper, of which the higher values indicate better performance.

B. Performance Analysis

Results analysis. First of all, it is important to analyze the performance of all IQA models within different λ ranges. We notice that with λ values closer to 0.5 (*i.e.,* the probability of causing strong visual confusion increases), nearly all metrics tend to perform worse. This indicates that the assessment of strong visual confusion is a difficult task for most models. As shown in Table I, λ can perform as the metric for confusion evaluation, and even acts better than MSE and PSNR, though the performance is still limited. Among classical IQA indexes, IW-SSIM and VIF show the top performances, which denotes that visual information weighting possibly helps the assessment of visual confusion. The improvements of introducing saliency into SSIM, FSIM, as well as GMSM demonstrate the importance of visual attention in visual confusion, which is worth further and deeper research. Surprisingly, the baseline deep features show good consistence with human perceiving on the entire dataset, though they are not well performed on either of the two sub-datasets. Unexpectedly, the widely used deep metrics LPIPS [37] and DISTS [62] perform even worse than the baseline methods, which may indicate that visual confusion is a different type of degradation compared to other distortions. Finally, our method gets relative better results and different backbone architectures show different optimization trends, which denotes that the feature extraction network is also important. Future studies on exploring different feature extraction methods are also needed.

Fig. 11 illustrates the performance evaluated by the new criteria on the CFIQA database. First, we observe that the proposed CFIQA model significantly outperforms other state-of-the-art models on *Different vs. Similar Analysis* and *Better vs. Worse Analysis* by a large margin. Furthermore, we notice the AUC values of the CFIQA metric on the *Better vs. Worse* classification task are higher than the *Different vs. Similar* classification task, which indicates that the *Different vs. Similar* classification is a more hard task and there is still room for improvement in this classification task.

Impact of different components. We further verify the impact of each component in our method. The analysis is conducted based on the backbone of VGG-16 and the results are shown in Table II. We first remove all components including channel attention, spatial attention and projection components (two-layer MLP), and only regress the weighting layers for feature fusion. The results shown in the first row of Table II demonstrate that the performance of this method is similar to the baseline method in Table I. Then we compare the impacts of channel attention and spatial attention modules for feature fusion. It can be observed that the spatial attention module contributes more to the final performance compared to the channel attention module. Furthermore, we compare

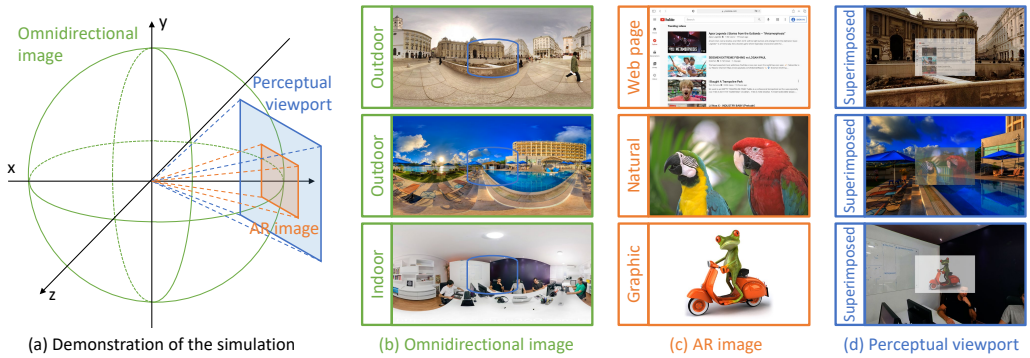


Fig. 12. The illustration of the AR simulation in VR environment. (a) The demonstration of the relationship between the omnidirectional image, the AR image, and the perceptual viewport image. (b) The omnidirectional images are used as the background scenes, which include outdoor and indoor scenarios. (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.

the contributions of two loss functions. It can be observed that ranking loss contributes most for constraining the optimization process. Though contributing less, the score regression loss still improves the performance.

VI. Extension Studies on Augmented Reality Image Quality Assessment (ARIQA)

In the above Study I and Study II, we have discussed a relatively more basic and more general problem, *i.e.*, visual confusion and its influence on the perceptual QoE. As aforementioned, visual confusion has significant influence on the QoE of human vision. However, the situation in the above studies is quite different with the real AR applications, which is mainly attributed to the fact that in actual AR applications, the FOV of the AR contents is usually smaller than the FOV of the real scenes [71]. Thus, we further conduct another subjective and objective IQA study towards evaluating the perceptual quality of AR contents in real-world AR applications.

A. Subjective ARIQA

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 [10], *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. 12 illustrates the methodology of the subjective experiment in this ARIQA study. First of all, 20 omnidirectional 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 may occur in AR applications. Three distortion types including image compression, image scaling, image contrast adjustment, are introduced as follows. (i) *JPEG compression* is a widely used method in image compression, and have been introduced into many quality assessment databases [72]. We set the quality level of the JPEG compression at the two levels with quality parameters 7 and 3. (ii) *Image scaling* is widely used in modern video streaming systems, where videos are often spatially downsampled prior to transmission, and then upscaled prior to display [73]. Such image scaling can also simulate the distortions introduced by various resolutions of AR devices. 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. (iii) *Image contrast adjustment* is also an important factor affecting the human visual perception and has been commonly introduced into natural IQA [74] and screen content IQA [72]. We also use the gamma transfer function [74] 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.

Since the visual confusion strongly influences the human visual perception as aforementioned, the superimposition degradation is also introduced. We design a program using Unity [75] to perform the experimental procedure, including stimuli display, data collection, *etc*. Fig. 13 demonstrates the user interface of the subjective ARIQA experiment. We first randomly match the 20 AR images and 20 omnidirectional

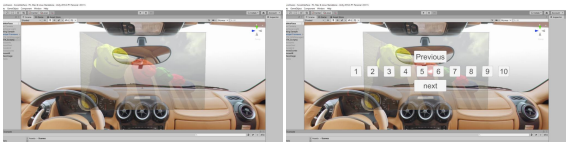


Fig. 13. Demonstration of the subjective experiment interface for ARIQA. Note that this subjective experiment is conducted in VR environment, however this demonstration is the screenshot from the desktop.

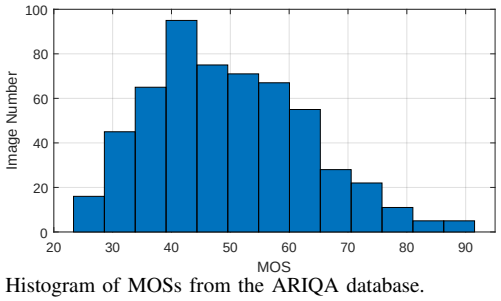


Fig. 14. Histogram of MOSs from the ARIQA database.

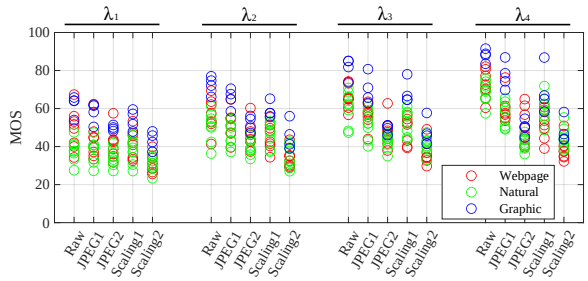


Fig. 15. Distribution of MOS values of raw images, JPEG compressed images, rescaled images superimposed on the omnidirectional backgrounds with different mixing values. The mixing values λ_1, λ_2, λ_3, λ_4 are equal to 0.26, 0.42, 0.58, 0.74, respectively.

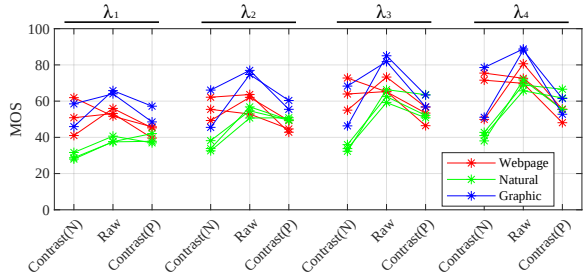


Fig. 16. Samples of MOS values of raw images, contrast adjusted images superimposed on the omnidirectional backgrounds with different mixing values. "N" denotes negative gamma transfer, "P" represents positive gamma transfer.

images 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:

$$I_S = \lambda \circ I_A + (1 - \lambda) \circ I_O \ , \tag{10}$$

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.*, four superimposed levels are introduced 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. 12 (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 by the perceptual viewport. Fig. 12 (b), (c) and (d) show the examples of the omnidirectional images, the AR images, and the perceptual viewport images, respectively.

A total of 23 subjects participate in the experiment, who are not included in the aforementioned CFIQA study. All subjects are recruited through standard procedures similar to that described in Section III-B. Before the formal test, each subject experiences a short training session where 28 stimuli are shown. The same distortion generation procedure is introduced for the training stimuli as for the test stimuli, and these training stimuli are not included in the test session. 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 [39]. We use HTC VIVE Pro Eye [76] as the HMD on account of its excellent graphics display technology and high precision tracking ability. During the formal test, all

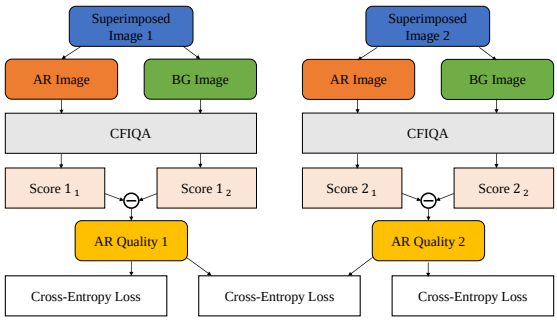


Fig. 17. The framework of the proposed ARIQA model.

560 experimental stimuli are displayed in a random order for each subject.

Subjective data processing and analysis. Similar to the procedure in Section III-C, we first process the collected subjective scores to obtain the MOSs. Only 1 subject is rejected, and each image is rated by 22 valid subjects. Among all scores given by the remaining valid subjects, about 3.21% of the total subjective evaluations are identified as outliers and removed. Fig. 14 plots the histogram of MOSs over the entire ARIQA database, showing a wide range of perceptual quality scores.

We further analyze the distribution of MOS values across different mixing values and various distortions. Fig. 15 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. Specifically, for the superimposition of raw AR images and background images, we find that graphic images provide better QoE than web page images, and web page images provide better QoE than natural images in general. It may reveal that relatively simple AR contents can provide better QoE than complex AR contents. 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. It may reveal that the superimposition degradation is a more influential quality factor compared to other distortions when the λ value is relatively small. However, it also means that the superimposition degradation can hide other distortions. Fig. 16 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.

B. Objective ARIQA Models

1) Benchmark models: As discussed in Section III and Section VI-A, the visual confusion may affect the human visual perception and may degrade the QoE of AR. However, whether the IQA metrics should consider the superimposed image, the AR image, and the background image together still needs to be discussed. Therefore, three benchmark variants are introduced for objective ARIQA. 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

TABLE III
PERFORMANCE OF THE THREE VARIANTS OF THE STATE-OF-THE-ART FR-IQA MODELS ON THE ARIQA DATABASE. 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↓	PWRC↑	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑
PSNR	0.2197	0.1485	0.2742	12.733	2.8868	0.0064 (−0.2133)	0.0027	0.0592	13.217	0.0299	0.3809 (+0.1612)	0.2662	0.4154	11.901	4.8525
NQM [48]	0.4101	0.2813	0.4268	11.974	5.9685	0.5348 (+0.1247)	0.3772	0.5550	11.014	7.7948	0.5588 (+0.1487)	0.4031	0.5867	10.677	7.6133
MS-SSIM [77]	0.6118	0.4414	0.6483	10.080	8.3464	0.6557 (+0.0439)	0.4778	0.6609	9.9366	9.0159	0.6660 (+0.0541)	0.4914	0.6741	9.6721	9.0949
SSIM [28]	0.5327	0.3799	0.5551	11.013	7.2294	0.5399 (+0.0072)	0.3797	0.5411	11.134	7.5044	0.6090 (+0.0763)	0.4430	0.6233	10.276	8.1620
IFC [49]	0.3539	0.2456	0.3294	12.501	5.6002	0.5121 (+0.1582)	0.3523	0.5105	11.385	7.2188	0.5090 (+0.1551)	0.3601	0.5217	11.172	7.0657
VIF [50]	0.5981	0.4273	0.6366	10.211	8.4390	0.6927 (+0.0946)	0.5009	0.6869	9.6218	9.5307	**0.7227** (+0.1245)	**0.5351**	**0.7222**	**9.2024**	**9.8505**
IW-MSE [51]	0.2287	0.1555	0.2966	12.644	3.0131	0.2406 (+0.0119)	0.1689	0.2956	12.648	3.7931	0.4126 (+0.1839)	0.2906	0.4586	11.693	5.7220
IW-PSNR [51]	0.2287	0.1555	0.2998	12.631	2.8814	0.2406 (+0.0119)	0.1689	0.2895	12.673	3.7414	0.3559 (+0.1272)	0.2574	0.4151	11.879	4.9256
IW-SSIM [51]	0.6431	0.4663	0.6532	10.026	8.8683	**0.7103** (+0.0672)	**0.5267**	**0.7100**	**9.3231**	**9.7540**	**0.7116** (+0.0685)	**0.5337**	**0.7201**	**9.0193**	**9.6804**
FSIM [29]	0.6323	0.4546	0.6723	9.8010	8.7569	0.6538 (+0.0215)	0.4716	0.6528	10.029	9.0221	0.6663 (+0.0340)	0.4865	0.6774	9.5764	9.3018
GSI [52]	0.4393	0.3046	0.5034	11.440	6.2671	0.3788 (−0.0605)	0.2606	0.3890	12.197	5.3120	0.4245 (−0.0147)	0.3056	0.4584	11.680	5.5771
GMSD [53]	0.6485	0.4718	0.6759	9.7575	8.8107	0.5947 (−0.0537)	0.4346	0.5959	10.633	7.7722	0.6730 (+0.0245)	0.4973	0.6801	9.5815	9.2104
GMSM [53]	0.6386	0.4628	0.6907	9.5745	8.7291	0.5863 (−0.0523)	0.4142	0.5923	10.667	8.0777	0.6294 (−0.0092)	0.4587	0.6422	10.064	8.5228
PAMSE [54]	0.2162	0.1458	0.2736	12.735	2.8281	0.0090 (−0.2072)	0.0048	0.0659	13.211	0.2350	0.3657 (+0.1495)	0.2558	0.4093	11.941	4.9560
LTG [55]	0.6592	0.4830	0.6826	9.6759	8.9866	0.6469 (−0.0123)	0.4742	0.6422	10.150	8.7329	0.6764 (+0.0172)	0.4998	0.6818	9.4727	9.3129
VSI [56]	0.5190	0.3691	0.5926	10.665	7.3012	0.6096 (+0.0906)	0.4318	0.6167	10.422	8.6713	0.6321 (+0.1131)	0.4590	0.6484	10.039	8.4581
LPIPS (Squeeze) [37]	0.5924	0.4326	0.6160	10.430	7.8301	0.6260 (+0.0336)	0.4450	0.6251	10.334	8.7742	0.6417 (+0.0494)	0.4693	0.6660	9.8086	8.7256
LPIPS (Alex) [37]	0.5870	0.4273	0.6314	10.267	7.7542	0.6306 (+0.0436)	0.4457	0.6352	10.226	8.8424	0.6626 (+0.0757)	0.4820	0.6767	9.6071	8.9045
LPIPS (VGG) [37]	0.5436	0.3828	0.5593	10.975	7.5461	0.6202 (+0.0766)	0.4426	0.6141	10.450	8.6116	0.6373 (+0.0936)	0.4606	0.6475	9.9848	8.4863
DISTS [62]	0.5011	0.3583	0.5280	11.244	7.4008	0.5112 (+0.0101)	0.3627	0.5528	11.033	6.6245	0.6334 (+0.1323)	0.4608	0.6580	9.7866	8.8983
Baseline (SqueezeNet)	0.5733	0.4166	0.6096	10.496	7.5174	0.6339 (+0.0606)	0.4570	0.6358	10.220	8.8444	0.6272 (+0.0539)	0.4573	0.6493	9.8747	8.7602
Baseline (AlexNet)	0.5273	0.3776	0.5814	10.772	6.9367	0.6450 (+0.1177)	0.4690	0.6578	9.9728	9.0323	0.6460 (+0.1187)	0.4768	0.6707	9.7499	8.9168
Baseline (VGG-16)	0.5541	0.3908	0.5706	10.873	7.7488	0.6368 (+0.0827)	0.4585	0.6372	10.204	8.8709	0.6587 (+0.1046)	0.4805	0.6622	9.8906	8.8234
Baseline+ (VGG-16)	0.6167	0.4454	0.5649	10.925	8.3908	0.6793 (+0.0626)	0.4979	0.6814	9.6902	9.4179	0.6815 (+0.0649)	0.5036	0.6896	9.4713	9.4016
Baseline (VGG-19)	0.5612	0.3981	0.5790	10.795	7.8228	0.6561 (+0.0949)	0.4750	0.6530	10.028	9.1239	0.6613 (+0.1001)	0.4838	0.6674	9.6720	9.1233
Baseline (ResNet-18)	0.5438	0.3892	0.5750	10.832	7.4822	0.6467 (+0.1029)	0.4678	0.6451	10.117	9.0764	0.6485 (+0.1047)	0.4779	0.6702	9.7504	9.0147
Baseline (ResNet-34)	0.5426	0.3862	0.5771	10.813	7.4832	0.6603 (+0.1177)	0.4782	0.6660	9.8765	9.2702	0.6710 (+0.1284)	0.4959	0.6903	9.4826	9.1111
Baseline (ResNet-50)	0.5753	0.4113	0.5977	10.615	7.9505	0.6510 (+0.0757)	0.4688	0.6464	10.102	9.1610	0.6619 (+0.0866)	0.4831	0.6736	9.7163	8.9478
CFIQA (VGG-19)	0.5002	0.3588	0.4881	11.556	6.5740	**0.7448** (+0.2446)	**0.5539**	**0.7436**	**8.8522**	**10.212**	**0.7511** (+0.2509)	**0.5648**	**0.7587**	**8.4569**	**10.080**
CFIQA (ResNet-34)	0.4061	0.2903	0.4405	11.886	5.3775	0.7024 (+0.2963)	0.5178	0.7092	9.3344	**9.6762**	**0.7092** (+0.3031)	**0.5280**	**0.7197**	**9.0977**	9.6462

TABLE IV
PERFORMANCE OF FOUR TRAINABLE MODELS.

Model \ Criteria	SRCC↑	KRCC↑	PLCC↑	RMSE↓	PWRC↑
LPIPS	0.7624	0.5756	0.7591	8.6935	10.936
CFIQA	0.7787	0.5863	0.7695	8.5484	11.125
ARIQA	0.7902	0.5967	0.7824	8.3295	11.314
ARIQA+	**0.8124**	**0.6184**	**0.8136**	**7.8018**	**11.576**

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: $I_A = \lambda \cdot I_{A_D}$, and $I_S = \lambda \cdot I_{A_D} + (1-\lambda) \cdot I_B$, respectively. Then, three FR-IQA benchmark variants used to calculate AR image quality 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 [78] 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:

$$Q_{\text{Type I}} = \text{FR}(I_A, I_{A_R}), \tag{11}$$

$$Q_{\text{Type II}} = \text{FR}(I_S, I_{A_R}), \tag{12}$$

$$Q_{\text{Type III}} = \text{SVR}(\text{FR}(I_S, I_{A_R}), \text{FR}(I_S, I_B)), \tag{13}$$

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

2) The proposed ARIQA model: Intuitively, using superimposed images as the perceptual distorted images, and considering their similarity with both AR references and background references may be more effective for evaluating the perceptual quality of the AR layers. Hence, as demonstrated in Figure 17, for better evaluating the perceptual quality of AR contents, we further improve the learning strategy of the CFIQA model to the ARIQA model by comparing the quality of two homologous superimposed images. Different from CFIQA, the goal for the ARIQA is to predict the perceptual quality of AR contents rather than both two views, therefore, the two output results of CFIQA are fused to predict the AR image quality. Considering the effectiveness of the training objectives of the LPIPS [37] and our CFIQA, during the training process, two pathways are introduced to ARIQA for comparing the perceptual quality of two different distorted images of one AR and background reference pair. Furthermore, we also improve the ARIQA model to the ARIQA+ model by incorporating the features from the edge detection model RCF net [66], which is similar to the way of the aforementioned CFIQA+.

C. Experimental Validation on the ARIQA Database

1) Experimental settings: In terms of our ARIQA database, 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 results of our proposed ARIQA model and benchmark models can be calculated accordingly as aforementioned.

Benchmark experiment setting. Besides the state-of-the-art FR-IQA metrics mentioned in Section IV-A, we also test the generalization ability of the CFIQA model trained on the CFIQA database on the ARIQA database. For the SVR experiment (*i.e.*, Type III mentioned above), we conduct a 100-fold cross validation experiment. For each fold, the ARIQA database is randomly split into a training dataset and a test dataset at a ratio of 4:1. The final results are calculated by averaging the test results of all 100 cross-validations.

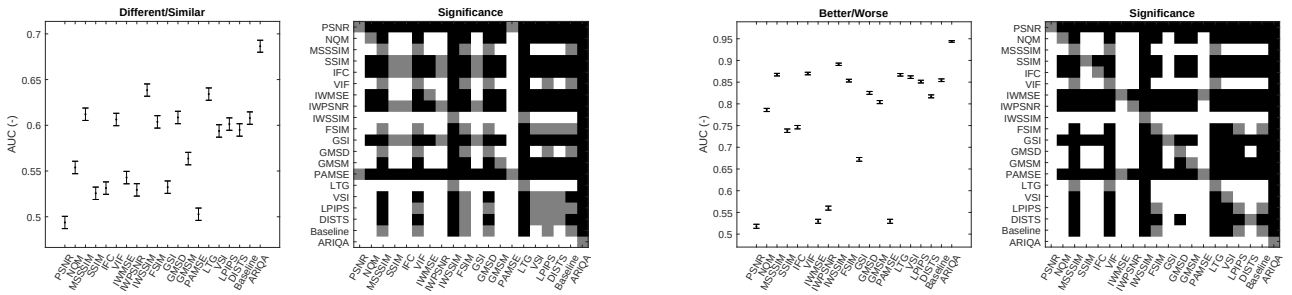


Fig. 18. New criteria performance of 19 state-of-art FR IQA models and the proposed metric on the ARIQA database. Left two figures are the different vs. similar ROC analysis results. Right two figures are the better vs. worse analysis results. The black/white/gray squares in the significance figures have the same meaning with that in Fig. 11

Deep learning-based experiment setting. We conduct a five-fold cross-validation experiment for the proposed ARIQA model on the ARIQA database. 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 also 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. Note that the CFIQA model here is a modified version, since the original CFIQA aims to compare the similarity between two reference images for one superimposed image, while here we focus on comparing two superimposed images for one AR reference, which is more similar to the concept of LPIPS.

2) Performance analysis: Table III presents the performance of three types of benchmark variants derived from the state-of-the-art FR-IQA models on the ARIQA database. 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. Moreover, it can be observed that our proposed CFIQA models trained on the CFIQA database also achieve state-of-the-art performance on the ARIQA database. This demonstrates the good generalization ability of our proposed CFIQA models and illustrates that visual confusion is one of the most important factors affecting the perceptual quality of AR images.

Table IV 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. Fig. 18 illustrates the performance evaluated by the new criteria on the ARIQA database. We notice that the proposed ARIQA model significantly outperforms other state-of-the-art models on *Different vs. Similar Analysis* and *Better vs. Worse Analysis* by a large margin.

VII. Conclusion and Future Works

In this paper, we discuss several AR devices and applications (see Section II-A), and clarify the essential theory underlying AR, *i.e.,* visual confusion. A more general problem underlying AR QoE assessment is first proposed, which is evaluating the perceptual quality of superimposed images, *i.e.*, confusing image quality assessment. To this end, we build a confusing image quality assessment (CFIQA) database, and conduct subjective and objective image quality assessment studies based on it. A CFIQA model is also proposed for better evaluating the perceptual quality of visual confusion. The results show that without extra degradation, the visual confusion itself can significantly influence the perceptual quality of the superimposed images, and state-of-the-art FR-IQA metrics are not well performed on this problem, especially when the mixing value is closer to 0.5. Moreover, the proposed CFIQA model performs better on this task. Next, in order to better study the influence of visual confusion on the perceptual quality of AR images in real application scenarios, an ARIQA study is further conducted. An ARIQA database is constructed, and three benchmark model variants as well as a specifically designed ARIQA model are proposed. 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 benchmark models. We hope this work can help other researchers have a better understanding of the visual confusion effect underlying AR technology. There are many issues related to visual confusion or AR QoE assessment that need to be explored in the future. Several key aspects are discussed as follows.

More complex degradations and diverse digital contents. In this paper, on the basis of visual confusion degradation, we further incorporate three other distortion types. However, we may encounter more complex degradations [79], [80] in daily life due to the limitation of photographic apparatus, compression processing, transmission bandwidth, display devices, *etc.* The perceptual peculiarity of these more complex distortions may be different from their original characteristics when mixed with the visual confusion degradation. Moreover, besides 2D AR contents, 3D rendering may also produce visual confusion effect when the 3D virtual content is not fitting in the real-world environment. The perceptual quality of these more complex degradations and diverse digital contents under the visual confusion condition needs to be further studied.

More realistic simulation for real-world scenes and AR devices. It should be noted that the simulation used in our ARIQA study can also be improved. First of all, in the ARIQA study, we use omnidirectional images as background scenes, which cannot simulate real-world depth cues. It is significant to consider using omnidirectional stereoscopic images or virtual

3D scenes as background scenes to study the 3D visual confusion problem in future studies. However, the stereoscopic factor is hard to control, which should be carefully designed. Moreover, some AR devices may be equipped with dynamic dimming functions, such as Microsoft HoloLens [23] and Magic Leap [24] *etc*. It is also important to discuss this aspect when studying visual confusion effect for specific devices in future works. Finally, in real AR applications, the dynamic range of background scenes and AR contents may be different, since the background scenes are usually optically see-through and directly observed by users, while the dynamic range of AR contents is limited by the display. However, due to the limitation of the VR display, we cannot reproduce the dynamic range of real-wold scenes. Future works can also consider using real AR devices to study the influence of visual confusion on the QoE of AR.

Visual attention/saliency of visual confusion. As can be observed in Section V above, incorporating saliency as spatial attention into IQA metrics introduces significant performance improvement. However, in this work, we just apply the saliency map predicted by current methods on the reference images. We notice that state-of-the-art saliency prediction models fail to predict the saliency on the superimposed distorted images. Since the superimposition may change the original texture, the visual attention may be altered correspondingly. Future studies on predicting the saliency of confusing (superimposed) images are also needed [45]. Moreover, the relationship between saliency prediction and quality assessment under the visual confusion situation also needs more discussion. In our ARIQA study, we notice that clearer AR contents bring better perceptual quality for them. However, it does not mean that we should make the AR as clear as possible, since it may completely occlude the background view and may cause trouble or even danger. For see-through views, besides the perceptual quality metric, we suppose that the visibility metric may be a good way for evaluating it, which needs more research on it.

How to improve the QoE of AR. Based on the above discussions, we suppose that it is important to consider how to *carefully design* and *harmoniously display* the digital contents of AR to make the QoE of both the virtual world and the real world better, especially for different application scenarios and user requirements. We discuss some factors and present several recommendations that may be deserved to be studied to improve the QoE of AR in the future as follows. **(i)** FOV. The main difference between the CFIQA database and the ARIQA database is that the FOVs of two superimposed views in the ARIQA database are different. We notice that the superimposition of two views with different fields may help distinguish each other. Future works on how to appropriately design the FOV for specific applications may be helpful. **(ii)** 3D depth cues. The depth difference between AR contents and BG scenes can help distinguish two layers and may improve the perceptual quality of them. However, it should be noted that the depth difference may increase the risk of inattentional blindness [81], [82]. Thus it may be better to study the influence of this factor in both the quality problem (as discussed in this paper) and the saliency problem [45] together to give a trade-off solution. **(iii)** Similarity/correlation between AR contents and BG scenes. As shown in Fig. 1, if the background view or augmented view is blank, there is no visual confusion. However, this situation is unlikely to happen in real cases. A more reasonable solution is calculating the similarity/correlation between AR view and BG view, then looking for appropriate space to display AR contents or adjusting the brightness/contrast/color of AR contents [13]. Our proposed method can be used in this situation. **(iv)** Displaying AR contents considering scenarios. It should be noted that different scenarios (*e.g.,* moving, talking, relaxing, *etc.*) may need different display solutions. This is a human-centric problem, which may need to combine other computer vision techniques from egocentric problems [83] for AR displaying.

Other computer vision tasks and applications. The directly related CV tasks to this work are blind source separation (BSS) [44], [84] and its sub-tasks, such as image reflection removal [44] *etc*. BSS aims at separating source signals/images from a set of mixed ones, which has been widely explored recently [42], [44]. Research on FR or NR IQA metrics of visual confusion may contribute to the evaluation of these problems. Moreover, besides the AR, the visual confusion may also appear in other display technologies, such as projector and transparent display screens [85]. Future work on NR-IQA metrics of visual confusion may help assess the QoE of these devices.

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. Guo, V. Vidal, I. Cheng, A. Basu, A. Baskurt, and G. Lavoue, "Subjective and objective visual quality assessment of textured 3d meshes," *ACM Transactions on Applied Perception (TAP)*, vol. 14, no. 2, pp. 1–20, 2016.

[4] H. Su, Z. Duanmu, W. Liu, Q. Liu, and Z. Wang, "Perceptual quality assessment of 3d point clouds," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3182–3186.

[5] E. Alexiou, T. Ebrahimi, M. V. Bernardo, M. Pereira, A. Pinheiro, L. A. D. S. Cruz, C. Duarte, L. G. Dmitrovic, E. Dumic, D. Matkovics *et al.*, "Point cloud subjective evaluation methodology based on 2d rendering," in *Proceedings of the International Conference on Quality of Multimedia Experience (QoMEX)*, 2018, pp. 1–6.

[6] 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.

[7] 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.

[8] E. Alexiou, E. Upenik, and T. Ebrahimi, "Towards subjective quality assessment of point cloud imaging in augmented reality," in *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2017, pp. 1–6.

[9] L. Zhang, H. Dong, and A. El Saddik, "Towards a qoe model to evaluate holographic augmented reality devices," *IEEE MultiMedia*, vol. 26, no. 2, pp. 21–32, 2018.

[10] 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.

[11] 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.

[12] 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.

[13] E. Ahn, S. Lee, and G. J. Kim, "Real-time adjustment of contrast saliency for improved information visibility in mobile augmented reality," *Virtual Reality*, vol. 22, no. 3, pp. 245–262, 2018.

[14] D. Kalkofen, E. Veas, S. Zollmann, M. Steinberger, and D. Schmalstieg, "Adaptive ghosted views for augmented reality," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2013, pp. 1–9.

[15] Double vision. [Online]. Available: https://www.drmilesburke.com/eye-conditions/eye-alignment-strabismus/double-vision.html

[16] J. R. Economides, D. L. Adams, and J. C. Horton, "Perception via the deviated eye in strabismus," *Journal of Neuroscience*, vol. 32, no. 30, pp. 10 286–10 295, 2012.

[17] L. A. Kelley and J. L. Kelley, "Animal visual illusion and confusion: the importance of a perceptual perspective," *Behavioral Ecology*, vol. 25, no. 3, pp. 450–463, 2014.

[18] H. Apfelbaum and E. Peli, "Tunnel vision prismatic field expansion: challenges and requirements," *Translational vision science & technology*, vol. 4, no. 6, pp. 8–8, 2015.

[19] R. Blake and N. K. Logothetis, "Visual competition," *Nature Reviews Neuroscience*, vol. 3, no. 1, pp. 13–21, 2002.

[20] Google glass. [Online]. Available: https://www.google.com/glass

[21] Vuzix m400. [Online]. Available: https://www.vuzix.com/products/m400-smart-glasses

[22] R. P. O'Shea, A. Parker, D. La Rooy, and D. Alais, "Monocular rivalry exhibits three hallmarks of binocular rivalry: Evidence for common processes," *Vision research*, vol. 49, no. 7, pp. 671–681, 2009.

[23] Hololens. [Online]. Available: https://www.microsoft.com/hololens

[24] Magic leap. [Online]. Available: https://www.magicleap.com

[25] Epson moverio augmented reality smart glasses. [Online]. Available: https://epson.com/moverio-augmented-reality

[26] S. Wang, K. Ma, H. Yeganeh, Z. Wang, and W. Lin, "A patch-structure representation method for quality assessment of contrast changed images," *IEEE Signal Processing Letters*, vol. 22, no. 12, pp. 2387–2390, 2015.

[27] X. Yu, C. G. Bampis, P. Gupta, and A. C. Bovik, "Predicting the quality of images compressed after distortion in two steps," *IEEE Transactions on Image Processing (TIP)*, vol. 28, no. 12, pp. 5757–5770, 2019.

[28] 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.

[29] 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.

[30] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2012.

[31] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia (TMM)*, vol. 20, no. 8, pp. 2049–2062, 2017.

[32] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia (TMM)*, vol. 17, no. 1, pp. 50–63, 2014.

[33] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *Proceedings of the IEEE international conference on image processing (ICIP)*, 2015, pp. 2791–2795.

[34] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 1, pp. 206–219, 2017.

[35] Y. Yuan, Q. Guo, and X. Lu, "Image quality assessment: a sparse learning way," *Neurocomputing*, vol. 159, pp. 227–241, 2015.

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European conference on computer vision (ECCV)*, 2016, pp. 694–711.

[37] 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.

[38] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu, "Ivqad 2017: An immersive video quality assessment database," in *Proceedings of the International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2017, pp. 1–5.

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

[40] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, "Mc360iqa: a multi-channel cnn for blind 360-degree image quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 64–77, 2019.

[41] W. Zhou, J. Xu, Q. Jiang, and Z. Chen, "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 (TCSVT)*, 2021.

[42] Z. Zou, S. Lei, T. Shi, Z. Shi, and J. Ye, "Deep adversarial decomposition: A unified framework for separating superimposed images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 806–12 816.

[43] H. Duan, X. Min, W. Shen, and G. Zhai, "A unified two-stage model for separating superimposed images," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2065–2069.

[44] H. Duan, W. Shen, X. Min, Y. Tian, J.-H. Jung, X. Yang, and G. Zhai, "Develop then rival: A human vision-inspired framework for superimposed image decomposition," *IEEE Transactions on Multimedia (TMM)*, 2022.

[45] H. Duan, W. Shen, X. Min, D. Tu, J. Li, and G. Zhai, "Saliency in augmented reality," in *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022, p. 6549–6558.

[46] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, no. 2, pp. 303–338, 2010.

[47] B. Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 500–13, 2012.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6629–6640.

[58] 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.

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

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

[61] 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.

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

[63] D. L. Yamins and J. J. DiCarlo, "Using goal-driven deep learning models to understand sensory cortex," *Nature Neuroscience*, vol. 19, no. 3, pp. 356–365, 2016.

[64] 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.

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

[66] 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.

[67] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[68] Q. Wu, H. Li, F. Meng, and K. N. Ngan, "A perceptually weighted rank correlation indicator for objective image quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 27, no. 5, pp. 2499–2513, 2018.

[69] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Proceedings of the IEEE International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[70] L. Krasula, P. Le Callet, K. Fliegel, and M. Klíma, "Quality assessment of sharpened images: Challenges, methodology, and objective metrics," *IEEE Transactions on Image Processing (TIP)*, vol. 26, no. 3, pp. 1496–1508, 2017.

[71] E. Kruijff, J. E. Swan, and S. Feiner, "Perceptual issues in augmented reality revisited," in *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2010, pp. 3–12.

[72] H. Yang, Y. Fang, and W. Lin, "Perceptual quality assessment of screen content images," *IEEE Transactions on Image Processing (TIP)*, vol. 24, no. 11, pp. 4408–4421, 2015.

[73] X. Min, G. Zhai, J. Zhou, M. C. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 6054–6068, 2020.

[74] 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.

[75] Unity. [Online]. Available: https://unity.com/

[76] Htc vive pro eye. [Online]. Available: https://www.vive.com/us/product/vive-pro-eye/overview/

[77] 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.

[78] 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.

[79] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal processing: Image communication*, vol. 30, pp. 57–77, 2015.

[80] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 4041–4056, 2020.

[81] A. Mack, "Inattentional blindness: Looking without seeing," *Current directions in psychological science*, vol. 12, no. 5, pp. 180–184, 2003.

[82] Y. Wang, Y. Wu, C. Chen, B. Wu, S. Ma, D. Wang, H. Li, and Z. Yang, "Inattentional blindness in augmented reality head-up display-assisted driving," *International Journal of Human–Computer Interaction*, pp. 1–14, 2021.

[83] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," *arXiv preprint arXiv:2110.07058*, vol. 3, 2021.

[84] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[85] Luxlabs. [Online]. Available: https://luxlabsdisplays.com/

Huiyu Duan received the B.E. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. From Sept. 2019 to Sept. 2020, he was a visiting Ph.D. student at the Schepens Eye Research Institute, Harvard Medical School, Boston, USA. His research interests include perceptual quality assessment and extended reality (XR).

Xiongkuo Min received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018, where he is currently a tenure-track Associate Professor with the Institute of Image Communication and Network Engineering. His research interests include image/video/audio quality assessment, quality of experience, visual attention modeling, extended reality, and multimodal signal processing.

Yucheng Zhu received the B.E. degree from the Shanghai Jiao Tong University, Shanghai, China, in 2015, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2021. He is currently a Post-Doctoral Fellow with Shanghai Jiao Tong University. His research interests include visual quality assessment, visual attention modeling and perceptual signal processing.

Guangtao Zhai (SM'19) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009, where he is currently a Research Professor with the Institute of Image Communication and Information Processing. He received the Award of National Excellent Ph.D. Thesis from the Ministry of Education of China in 2012. His research interests include multimedia signal processing and perceptual signal processing.

Xiaokang Yang (M'00-SM'04-F'19) received the B.S. degree from Xiamen University, Xiamen, China, in 1994, the M.S. degree from the Chinese Academy of Sciences, Shanghai, China, in 1997, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, in 2000. His current research interests include image processing and communication, computer vision, and machine learning. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and a Senior Associate Editor of the IEEE SIGNAL PROCESSING LETTERS.

Patrick Le Callet (F'19) received the M.Sc. and Ph.D. degrees in image processing from the Ecole Polytechnique de 1'Université de Nantes. He was an Assistant Professor from 1997 to 1999 and a full time Lecturer from 1999 to 2003 with the Department of Electrical Engineering, Technical Institute of the University of Nantes. He led the Image and Video Communication Laboratory, CNRS IRCCyN, from 2006 to 2016, and was one of the five members of the Steering Board of CNRS, from 2013 to 2016. Since 2015, he has been the Scientific Director of the cluster Ouest Industries Cratives, a five-year program gathering over ten institutions (including three universities). Since 2017, he has been one of the seven members of the Steering Board of the CNRS LS2N Laboratory (450 researchers), as a Representative of Polytech Nantes. Since 2019, he has been the fellow of IEEE. He is mostly involved in research dealing with the application of human vision modeling in image and video processing.